Methodische Anmerkungen zur Auswertung der WWW-Log-Dateien des Servers www.gesis.org

Mell, Wolf-Dieter

Veröffentlichungsversion / Published Version
Arbeitspapier / working paper

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Mell, W.-D. (2002). *Methodische Anmerkungen zur Auswertung der WWW-Log-Dateien des Servers www.gesis.org.* (IZ-Arbeitsbericht, 26). Bonn: Informationszentrum Sozialwissenschaften. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-50749-5

GESIS

IZ InformationsZentrum Sozialwissenschaften

IZ-Arbeitsbericht Nr. 26

Methodische Anmerkungen zur Auswertung der WWW-Log-Dateien des Servers www.gesis.org

Wolf-Dieter Mell

Juli 2002

InformationsZentrum
Sozialwissenschaften

Lennéstraße 30
D-53113 Bonn
Tel.: 0228/2281-0
Fax.: 0228/2281-120
email: mell@bonn.iz-soz.de
Internet: http://www.gesis.org

ISSN:	1431-6943
Herausgeber:	Informationszentrum Sozialwissenschaften der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e.V. (ASI)
Druck u. Vertrieb:	Informationszentrum Sozialwissenschaften, Bonn Printed in Germany

Das IZ ist Mitglied der Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen e.V. (GESIS), einer Einrichtung der Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz (WGL)

Inhalt

1 Randbedingungen

1.1 Vorbemerkung

Mit Datum vom 12.4.2001 wurde vom Autor eine "Studie zur Internet-Nutzung ausgewählter Anwendungen auf dem Server www.bonn.iz-soz.de im Jahr 2000" erstellt und auf Abteilungsleiter-Ebene im IZ präsentiert. Die Auswertungen der WWW-Log-Dateien erfolgte mit Hilfe des Tools WebSuxess. Die Ergebnisse wurden vom Autor fachlich interpretiert und mit den Umsatzzahlen für die konventionellen Produkte des IZ verglichen.

Die Diskussion dieser Studie ergab - abgesehen von unterschiedlichen Auffassungen bei der Bewertungen der Ergebnisse - eine Reihe von Schwachpunkten, möglichen Fehlerquellen und Lücken in der Analyse und der Aggregation der Log-Dateien.

Es wurde deshalb vereinbart, die Datenanalyse und die Datenzusammenfassung methodisch und technisch systematisch zu überprüfen und die Ergebnisse - ggf. mit methodischen Vorschlägen für eine regelmäßige Wiederholung der Auswertung - zu gegebener Zeit erneut vorzulegen.

Der vorliegende Bericht wurde als Entwurf im Januar 2002 vorgelegt und IZ-intern diskutiert. Auf der Grundlage der Vorschläge in diesem Bericht und der Besprechungsergebnisse wurde das Produktionsverfahren implementiert, der Bericht entsprechend ergänzt und im Juli 2002 veröffentlicht.

1.2 Ziel der Untersuchung

Die vorliegende Untersuchung hat drei Ziele:

- Es soll grundsätzlich geprüft werden, mit welchen Aspekten der Log-Dateien welche realistischen Aussagen über die tatsächliche Nutzung des WWW-Angebotes gemacht werden können. Hierbei geht es weniger darum, innovative neue Interpretationsgesichtspunkte zu entwickeln, als vielmehr darum, Realitätsnähe und Verlässlichkeit der zukünftig operativ verwendeten Nutzungskennziffern sicher zu stellen.

- Es werden Empfehlungen benötigt, welche Nutzungszahlen als vergleichbare Kennziffern im nationalen und internationalen Vergleich der Nutzung von WWW-Angeboten für den GESIS-Server verwendet werden sollen.

- Die einzelnen Abteilungen des IZ sind daran interessiert, Nutzungskennzahlen ihrer Angebotsteile als regelmäßiges feedback zu erhalten.

1.3 Datenbasis und Tools

Die Auswertungen in dieser Untersuchung basieren auf den Log-Dateien des Servers

WWW.GESIS.ORG (193.175.239.100),
Betriebssystem: Windows 2000 AS,
Serversoftware: IIS 5.0,
Zeitraum: 31.1.2001 - 31.12.2001,
ergänzende Auswertungen: 1.1.2002 - 13.6.2002

Die Auswertung erfolgt einerseits mit dem Analysetool WebSuxess 4, andererseits mit Hilfe von Filterskripten, mit denen Teilmengen der Log-Datei erzeugt werden.

Die Filterprozeduren sind so konzipiert, dass die Filterbedingungen (z.B. URL-Listen oder Listen mit IP-Adressen) aus ASCII-Input-Dateien eingelesen werden und dort auch gepflegt werden können.

Die Output-Dateien der Filterprozeduren sind formatkompatibel mit der ursprünglichen Log-Datei und können als Input für WebSuxess verwendet werden. Alternativ können sie durch ergänzende Prozeduren umformatiert werden, z.B. als Input für Excel-Auswertungen.

Als Programmiersprache für die Filterprozeduren wird AWK in einer Batch-Umgebung verwendet. Dieses Verfahren ermöglicht ein schnelles und einfaches Prototyping, ist allerdings unter Performance-Gesichtspunkten für den Produktionseinsatz weniger geeignet.

2 Probleme und Lösungsoptionen

2.1 Vorbemerkung

Es ist üblich und notwendig, die Nutzung der WWW-Dienste durch Kennzahlen zu quantifizieren.

Hierzu wird häufig die Anzahl der Hits verwendet, da dies die eindruckvollsten Zahlen liefert. Diese Kennzahl ist deswegen problematisch, weil sie alle Teilaufrufe einschließlich der Bilder, Ikons, Navigationshilfen und Applets enthält. Es ist bei seriösen Anbietern daher inzwischen üblich, als Kennzahl die PageViews auf die Inhalts-Seiten zu verwenden. Unklar bleibt bei veröffentlichten Zahlen i.d.R., in welchem Umfang eigene Zugriffe für Pflegemaßnahmen, Suchmaschinenzugriffe, regelmäßige Proxy-Abfragen oder regelmäßige "Verfügbar-Prüfungen" (z.B. durch Leitstände) in diesen Zahlen enthalten sind und welche Größenordnung die dadurch entstehenden Verzerrungen haben.

Wie sich gezeigt hat, schützt auch die Wahl eines guten Auswertungs-Tools nicht vor Analysefehlern, da relevante Gesichtspunkte (z.B. Filtern von Suchmaschinen) oft nur mit großem Aufwand oder gar nicht berücksichtigt werden können. Die Auswertungen der Studie vom 12.4.2001 basierten ausschließlich auf den Ergebnisdaten von WebSuxess. Wie sich bei nachträglicher, detaillierter Untersuchung herausstellte, wurden dabei einige Fehlerquellen übersehen oder nicht angemessen berücksichtigt, u.a.:

- Die Zugriffe der Suchmaschinen sind bei WebSuxess grundsätzlich in allen Seiten-Zugriffszahlen enthalten, sie werden nicht - wie irrtümlich angenommen - gefiltert. Darüber hinaus ist in den Suchmaschinen-Listen von WebSuxess nur ein Teil der tatsächlich aktiven Suchmaschinen enthalten.

- Das Konzept des "Mehrfach-Besuchers" (als Pendant eines "Kunden") auf der Basis der Identitäts-Erkennung mit Hilfe der IP-Adresse wird zunehmend problematisch, da sowohl der Anteil der Proxy-Server als auch die Anzahl von Zugriffen aus Provider-Netzen mit temporärer IP-Adressen-Zuweisung (DHCP) steigt und nicht mehr vernachlässigt werden kann.

- Die Summierung der Zugriffszahlen von Page-Gruppen - z.B. kompletten Unterverzeichnissen - als akkumulierte Nutzungskennzahl erweist sich als problematisch, da diese Summen nur beschränkt miteinander vergleichbar sind.

Im folgenden sollen sowohl die prinzipiellen Eigenschaften der Auswertungsdaten als auch die speziellen Vor- und Nachteile ausgewählter Auswertungs-Tools systematisch behandelt werden.

2.2 Prinzipielle Interpretations-Probleme

2.2.1 Terminologie

Zur Analyse von WWW-Log-Daten wird üblicherweise folgende Terminologie verwendet:

Hits: Gesamtzahl der einzelnen Zugriffe inkl. aller Bilder und Ikons.

PageViews: Anzahl der Seitenzugriffe mit herausgefilterten Bildern, Ikons, Applets etc.
Dieses Maß wird im folgenden als relevantes Maß für den Zugriff auf Einzel-Informationen verwendet

Besucher: Der durch die IP-Adresse identifizierte individuelle Nutzer des Angebotes. Sowohl bei Internet-Zugang über Telefon/ISDN-Gateways (z.B. RAS) als auch bei der IP-Zuteilung per DHCP ist eine verlässliche Zuordnung von Personen zu IP-Adressen nicht möglich. Ein weiterer Aspekt ist der zunehmende Einsatz von Proxy-Servern aus Sicherheits- oder Performance-Gründen, wodurch ein verlässlicher Rückschluss auf Anzahl oder Identität der Endnutzer-Zugriffe zusätzlich erschwert wird.

Einmalige Besucher: Anzahl der IP-Adressen, für die im Auswertungszeitraum (im vorliegend Fall das Jahr 2001) nur eine Visit beobachtet wurde. Im Gegensatz dazu: Mehrfach-Besucher (= Anzahl Besucher minus einmalige Besucher), eine allerdings im Hinblick auf DHCP und Proxy problematische Kennzahl.

Visits: Zusammenfassung der zeitlich zusammenhängenden Aktivitäten eines "Besuchers" auf dem Server zu einer kumulierten Kennzahl, die als zusammenhängender Besuch auf einer Web-Site interpretiert werden kann.
Die Anzahl der von Mehrfach-Besuchern durchgeführten Visits ergibt sich aus der Gesamtzahl der Visits abzüglich der Anzahl der Einmal-Besucher. Die Zahl der Mehrfach-

Besucher-Visits pro Mehrfach-Besucher könnte ggf. als Kennzahl für die Intensität der Nutzung durch aktive Kunden gewertet werden, sofern DHCP- und Proxy-Effekte ausreichend sicher ausgeschaltet werden können.

2.2.2 Felder der Log-Datei

IIS 5.0 erzeugt eine Log-Datei der WWW-Zugriffe mit einem Record pro Zugriff und folgender, durch Leerstellen getrennter Feldstruktur:

Nr.	Feldbezeichnung	Beispiel	Bemerkung
1	date	2001-01-31	Datum
2	time	17:04:42	Uhrzeit des Zugriffs
3	c-ip	193.175.239.113	IP-Adresse des Client
4	cs-username	-	
5	cs-method	get	Befehlstyp
6	cs-uri-stem	/information/index.htm	URL auf dem Server
7	cs-uri-query	-	Abfrage-Parameter
8	sc-status	200	HTTP-Statuscode
9	sc-bytes	17874	übertragene Bytes
10	cs(User-Agent)	mozilla/4.7+[de]+(win98;+i)	Browser des Client
11	cs(Referer)	http://www.gesis.org/index.htm	vorherige URL

2.2.3 Identifikation des Besuchers

2.2.3.1 Problem: Suchmaschinen

Das Verfahren WebSuxess liefert mit den Dateien search2.dat und spiders.dat eine Liste von Suchmaschinen mit ihren Robotern. Diese Liste ist in dem uns zur Verfügung stehenden Stand nicht annähernd vollständig.

Es wurde daher anhand der Web-Zugriffe empirisch eine eigene Liste der IP-Adressen von Suchmaschinen aufgebaut. Die Selektion erfolgte nach folgenden Gesichtspunkten:

- Die Suchmaschinen arbeiten i.d.R. mit Gruppen von Robotern, deren IP-Adressen in allen beobachteten Fällen pro Suchmaschine innerhalb eines oder mehrerer 256-Knoten-Subnetze (Subnetz-Maske: 255.255.255.0) liegen. D.h.: Mehrere zusammengehörige Roboter einer Suchmaschine werden dadurch identifiziert, dass ihre IP-Adressen in den führenden 3 Bytes übereinstimmen.

- Die DNS-Namen der Roboter einer Suchmaschine haben typische Bezeichnungen:
  Domänen-Name = Firmen-Identifikation,

Host-Name = "spider", "search", "robot", "crawl" o.ä.
mit einer einheitlichen Systematik der Hostnamen innerhalb eines Subnetzes (z.B. spider01, spider02 etc.). Auf eine Suchmaschine wird also dann geschlossen, wenn mehrere Besucher mit dem gleichen Subnetz und dieser DNS-Namen-Systematik beobachtet werden.

- Wurden mehrere Suchmaschinen-Roboter in einem Subnetz identifiziert, so wurde kein Fall beobachtet, in dem ein Host aus diesem Subnetz mit einer von der DNS-Namen-Systematik der übrigen Hosts in diesem Subnetz abweichenden Benennung - als Hinweis auf eine andere Nutzung - auf den Web-Server zugegriffen hat.
  Hieraus wird geschlossen, dass ohne Auswirkungen auf die Besucher-Statistik in aller Regel das komplette Subnetz als "Suchmaschinen-Filter" verwendet werden kann.

- Suchmaschinen-Roboter fallen dadurch auf, dass für mindestens einen Roboter aus einem Suchmaschinen-Subnetz eine erheblich über dem Durchschnitt liegende Anzahl von PageViews angezeigt wird.

- Suchmaschinen verwenden spezielle Browser-Software (s.o. cs(User-Agent)), z.B. googlebot, scooter, fast-webcrawler u.ä.

Neben den rund 300 expliziten Suchroboter-Adressen aus den Listen von WebSuxess wurden mit dieser Systematik weitere rund 20 Suchmaschinen-Subnetze identifiziert und mit je 255 Adressen dem Suchmaschinen-Filter hinzugefügt.

Bei der Ermittlung der operativen Nutzungszahlen werden im folgenden Suchmaschinen-Zugriffe nicht als Besucher-Zugriffe gezählt.

2.2.3.2 Problem: Proxy

Proxy-Server sind spezielle WWW-Client-Systeme, die stellvertretend für die Endbenutzer einer Organisation mit den WWW-Servern kommunizieren und die übertragenen Seiten zwischenspeichern. Hierdurch können einerseits die übertragenen Seiten auf unerlaubte Inhalte, Viren etc. überprüft werden, andererseits Seiten, die auf dem Proxy-Server gespeichert wurden, anschließend von anderen Benutzern abgerufen werden, ohne dass erneut eine Verbindung zum zugehörigen WWW-Server hergestellt werden muss.

Insbesondere aus Sicherheitsgründen werden bei größeren Organisationen (Universitäten, Provider etc.) Proxy-Server in zunehmendem Umfang eingesetzt.

Proxy-Server nutzen in vielen Fällen eine spezielle Browser-Software (cs(User-Agent) s.o.) und fallen in den Nutzer-Statistiken nach Ausfiltern der Suchmaschinen durch ungewöhnlich hohe PageView-Zahlen pro Einzel-Host sowie in vielen Fällen durch spezielle, sprechende DNS-Namen mit Anteilen wie "cache", "proxy" etc. auf.

Je nach Konfiguration prüfen Proxy-Server bei einer erneuten Anforderung einer bereits gespeicherten Seite durch einen Benutzer

- entweder bei jeder Seitenanforderung, ob diese Seite auf dem zugehörigen WWW-Server gegenüber der gespeicherten Seite verändert wurde, in diesem Fall wird die Seite erneut geladen, anderenfalls wird die gespeicherte Seite dem Benutzer präsentiert.

- oder ob die letzte Überprüfung älter als z.B. 1 Tag ist, in diesem Fall wird die Aktualität auf dem WWW-Server überprüft, anderenfalls wird die gespeicherte Seite ohne erneuten WWW-Zugriff dem Benutzer zur Verfügung gestellt.

- In Einzelfällen sind Proxy-Server so konfiguriert, dass sie ausgewählte Seiten regelmäßig von den WWW-Servern aktualisieren, ohne dass hierzu Benutzeranforderungen als Trigger erforderlich sind.

Hinweis:

Alle relevanten Browser auf Endbenutzer-Systemen (Netscape, IExplorer etc.) sind standardmäßig so konfiguriert, dass sie genau wie ein Proxy die übertragenen Seiten in einem Cache zwischenspeichern und bei erneutem Zugriff des Anwenders auf diese Seite diese zunächst aus dem Cache präsentieren, ohne erneut auf dem WWW-Server zuzugreifen. Die "Lebensdauer" der Cache-Seiten kann konfiguriert werden, z.B. "nur während der laufenden Sitzung" oder eine Anzahl von Tagen.

Der Vergleich einer gespeicherten Seite mit dem Original auf einen WWW-Server erzeugt spezielle, im WWW-Log dokumentierte Statuscodes (cs-status s.o.).

Aus der Sicht der Besucher-Analyse haben Proxy-Server folgende Auswirkungen:

- Die Einzelbenutzer der Organisation und ihr Zugriffs-Verhalten sind nicht mehr erkennbar, da sie im WWW-Log durch die Aktionen des Proxy-Server ersetzt werden.

- Szenario 1: Viele Benutzer der Organisation nutzen "gleichzeitig" (z.B. innerhalb des gleichen Tages) die gleichen WWW-Seiten:
  a) Die Anzahl der im WWW-Log angezeigten PageViews ist erheblich geringer, als die Anzahl der tatsächlichen Nutzungen,
  b) Die "vielen" realen Benutzer werden im WWW-Log als 1 "Mehrfach-Besucher" abgebildet.

- Szenario 2: Wenige Benutzer der Organisation nutzen "selten und nicht gleichzeitig" die gleichen WWW-Seiten:
  a) Die Anzahl der im WWW-Log angezeigten PageViews entspricht in etwa der tatsächlichen Nutzung, da der Proxy bei Seitenanforderungen (fast) jedes Mal erneut auf den WWW-Server zugreift, um seinen Cache zu aktualisieren.
  b) Mehrere unterschiedliche Benutzer werden als 1 "Mehrfach-Besucher" abgebildet.

Eine erste grobe Prüfung der Nutzungszahlen lässt vermuten, dass mindestens 20% der PageViews von Proxy-Servern stammen.

2.2.3.3 Problem: DHCP und Adress Translation

Jeder Arbeitsplatz benötigt für einen Zugang zum Internet eine "gültige", international nur einmal vorhandene, vom Service-Provider zugewiesene und in die internationalen Router-Netze eingetragene IP-Adresse. Diese ist die "Hausnummer" über die jeder Internet-Knoten eindeutig adressiert werden kann.

Es ist in vielen größeren Organisationen, sowie bei remote-Zugang über Service-Provider üblich, den einzelnen Arbeitsplätzen keine statischen IP-Adressen fest zuzuordnen, sondern bei jeder Anmeldung dem Arbeitsplatz eine "beliebige" IP-Adresse aus einem Adress-Pool aktuell zuzuweisen.

Das Verfahren heißt DHCP und hat u.a. den Vorteil, dass insbesondere bei einer großen Anzahl von Kunden der verfügbare IP-Adressen-Pool rationeller

genutzt werden kann, da nur aktive Arbeitsplätze eine IP-Adresse zugewiesen bekommen.

Ein anderes Verfahren, bei dem ebenfalls dynamisch IP-Adressen zugeordnet werden, ist die Adress Translation. Hierbei wird einem Arbeitsplatz innerhalb eines geschlossenen Netzes durch einen Firewall/Gateway dann eine im Internet gültige IP-Adresse temporär zugewiesen, wenn dieser eine Kommunikationsanforderung mit einer Internet-Adresse an den Firewall/Gateway sendet. Der Zweck des Verfahrens ist es, einerseits knappe IP-Adressen wirtschaftlich zu verwalten, andererseits die Adressierung und damit den Zugang zu den Hosts innerhalb des geschlossenen Netzes nach außen zu verdecken.

Beide Verfahren haben aus der Sicht der Besucher-Analyse die gleiche Wirkung:

- Die gleiche IP-Adresse kann zu unterschiedlichen Zeiten unterschiedlichen Arbeitsplätzen zugewiesen sein.
- Ein Arbeitsplatz verwendet zu unterschiedlichen Zeiten unterschiedliche IP-Adressen.
- Die Anzahl der genutzten IP-Adressen ist pro Organisation, insbesondere bei Remote-Providern (T-Online, AOL etc.), i.d.R. erheblich geringer, als die Anzahl der potentiellen Arbeitsplätze.

Die Zählung der PageViews ist durch diese Verfahren nicht betroffen, es kann allerdings aus der IP-Adresse nicht mehr auf die Identität des Besuchers oder die Eigenschaft des "Mehrfach-Besuches des gleichen Besuchers" geschlossen werden.

IP-Adressen aus DHCP oder Adress Translation sind im WWW-Log typischerweise an den DNS-Namen der Provider (bn-online.de, t-online.de, aol.com etc.) sowie an Hostnamen mit laufender Numerierung (d123, p003 etc.) erkennbar.

Eine erste grobe Prüfung der DNS-Namen zu den IP-Adressen in der WWW-Log-Datei lässt vermuten, dass mindestens 30% der Besucher-IP-Adressen nicht fest einem Arbeitsplatz zugeordnet sind sondern temporär per DHCP o.ä. vergeben wurden.

2.2.3.4 Einmal-/Mehrfach-Besucher

Als "Mehrfach-Besucher" wird eine IP-Adresse bezeichnet, für die im Untersuchungszeitraum mehrere "Visits" festgestellt wurden.

Wie oben nachgewiesen, ist es wegen zunehmenden Einsatzes von Proxy- und DHCP-Technologie nicht (mehr) möglich, IP-Adressen zu "personalisieren" und auf individuelle Besucher abzubilden.

Wegen der Abbildung einer größeren Zahl von Arbeitsplätzen auf eine geringere Zahl von IP-Adressen durch die Proxy- und DHCP-Verfahren kann statistisch davon ausgegangen werden, dass

- die Anzahl der realen "Besucher-Individuen" größer ist als die Anzahl der "IP-Adressen-Besucher" (da sich mehrere Individuen nacheinander die gleiche Adresse teilen),
- die Anzahl der realen "Mehrfach-Besucher-Individuen" geringer ist als die Anzahl der "IP-Adressen-Mehrfach-Besuchern" (da die gleiche Adresse mehrfach nacheinander von unterschiedlichen Individuen benutzt wird).

Die Anzahl der im WWW-Log aufgezeigten Mehrfach-Besucher ist als Kennzahl damit eine Obergrenze für das Mehrfach-Interesses der Gesamtheit der Mitglieder von Organisationen oder Providern an den Inhalten des WWW-Servers, sofern Mehrfach-Zugriffe nicht durch Automaten generiert werden.

2.2.4 Hits, PageViews, Visits

Wie in Kap. 2.2.1 dargestellt, beziehen sich die Begriffe "Hits" und "PageViews" jeweils auf die Zugriffe auf einzelne Seiten. Als PageViews gelten dabei Zugriffe auf den Inhalt der Seiten, Zugriffe auf Ikons, Navigationshilfen, Applets etc. werden dazu herausgefiltert.

Die PageView-Zählungen für einzelne Seiten sind unabhängig davon, ob die komplette Log-Datei eines Zeitraumes oder gefilterte Ausschnitte aus der Log-Datei mit diesen Seiten des gleichen Zeitraumes ausgewertet werden.

Visits dagegen beziehen sich auf die Aktivitäten von IP-Adressen ("Besuchern") auf dem gesamten Server-Angebot, bzw. dem in der Log-Datei oder deren gefilterten Ausschnitt dargestellten Teil.

Eine Visit ist definiert als eine Aktivität, bei der eine beliebige Zahl von Zugriffen der gleichen IP-Adresse auf beliebige Seiten des WWW-Angebotes mit zeitlichen Abständen zwischen den einzelnen Zugriffen unterhalb einer konfigurierbaren Schwelle (z.B. 10 Minuten) beobachtet werden.

Da Visits aus der Log-Datei bzw. derem gefilterten Ausschnitt rekonstruiert werden, verliert dieser Begriff an Bedeutung, wenn die Auswertung sich auf gefilterte Teile der Log-Datei, z.B. Datensätze bestimmter Seiten konzentriert, da dann Sprünge der gleichen IP-Adresse in andere Teile des WWW-Angebotes nicht registriert werden können.

2.2.5 URLs, Befehlstyp, Statuscode

Der Log-Datensatz eines Zugriffs auf den Server enthält u.a. ein Feld mit der URL des Elementes, das vom Client aufgerufen wurde (s.o. cs-uri-stem). Die Protokollierung erfolgt ohne Protokollbezeichnung (z.B. http://) und ohne den Servernamen (z.B. WWW.GESIS.ORG). Dieses Feld eignet sich dazu, die inhaltlichen Seiten von besonderem Auswertungsinteresse (z.B. /information/themen/fokus/index.htm) zu filtern, um sie z.B. einer Detailanalyse zu unterziehen.

Anmerkung:

Es ist zu beachten, dass der Eintrag im URL-Feld einschließlich Groß/Kleinschreibung so erfolgt, wie vom Client - ggf. inkl. Schreibfehler - übertragen, sodass bei Auswertungen - einschließlich WebSuxess - damit gerechnet werden muss, die gleiche URL in mehreren Groß/Klein-Schreibweisen sowie mit fehlerhaften Zusätzen mehrfach vorzufinden. Für die vorliegenden Auswertungen werden deshalb alle Texte der Log-Datei auf Kleinschreibung standardisiert und Schreibfehler - soweit erkannt und soweit sie nicht durch Statuscodes identifiziert werden - herausgefiltert.

Der Befehlstyp (s.o. cs-method) kennzeichnet die Art der Anforderung an den Server. In der Log-Datei wurde in der weit überwiegenden Zahl der Fälle der Befehlstyp "get" gefunden, in Einzelfällen der Befehlstyp "head". Der Befehlstyp wird in den vorliegenden Auswertungen nicht berücksichtigt.

Der Statuscode (s.o. sc-status) kennzeichnet die Reaktion des Servers auf die Anforderung des Client. Es wurden in der Log-Datei u.a. beobachtet:

Statuscode	Bedeutung
200	OK, Seitenanforderung erfüllt
206	Partitial Content, Übermittlung eines Teil-Objektes

304	Not Modified, Antwort auf die Anfrage z.B. eines Proxy
404	Not found, Seite nicht verfügbar
406	None Acceptable, Client kann Antwort nicht annehmen.

Die Statuscodes 200, 206 und 304 werden als Indiz für erfolgreich bearbeitete Seitenabfragen gewertet.

2.2.6 Referrer - vorherige URL

Jeder Datensatz der Log-Datei enthält ein Feld, in dem die URL dokumentiert wird, von der aus auf das vorliegende Element gesprungen wurde (s.o. cs(Referer)). Diese Information wird vom Browser des Client geliefert, sofern dieser dazu technisch in der Lage ist und diese Funktion konfiguriert wurde.

Durch Verketten der Referrer-Informationen aus einer Visit können Hinweise auf das Arbeitsverhalten des Besuchers gewonnen werden. Die zeitlich erste Referrer-URL einer Visit ist häufig der Link - z.B. einer Suchmaschine - die auf das GESIS-Angebot verwiesen hat (s. Kap. 5.5.3).

Referrer-Analysen sind fachlich interessant, aber technisch sehr aufwendig und schwierig zu interpretieren.

Eine spezielle Untermenge der Refferrer-Auswertungen ist die Analyse von Links aus bekannten Web-Sites - z.B. Suchmaschinen - um die Wirksamkeit von Link- und/oder Werbemaßnahmen zu überprüfen.

2.2.7 IVW-/Rawena-Verfahren

Die IVW (Informationsgemeinschaft zur Feststellung der Verbreitung von Werbeträgern e.V.) ist ein Zusammenschluss verschiedener Verbände zur Kontrolle der Auflagenstärken von Zeitungen, Zeitschriften und seit 1997 auch Internetangeboten. Die IVW veröffentlicht jeden Monat die Zugriffszahlen der Online-Angebote ihrer Mitglieder.

Die Messung der Zugriffszahlen erfolgt einheitlich mit dem "Verfahren zur Messung von Online-Reichweite" (Rawena) der Firma Ecce Terram in Oldenburg.

Hierbei werden auf jeder HTML-Seite Zählimpulse in Form unsichtbarer Graphiken eingebaut, die von einer Zählkomponente bearbeitet und von einer Auswertungskomponente nach festgelegten Regeln ausgewertet werden.

Die in dieser Auswertung verwendete Definition von PageView entspricht dabei in etwa der Definition einer PageView in Rawenna.

2.3 Folgerungen

2.3.1 Anforderungen an die Ergebnisse der Auswertung

Insgesamt werden an die Ergebnisse der WWW-Log-Auswertungen aus Sicht des operativen Bedarfs im IZ folgende Anforderungen gestellt:

- Die Fachabteilungen erwarten wenige, aussagekräftige Kennzahlen für die Nutzung ihrer einzelnen Produkte.
- Zusätzlich werden Zeitreihen pro Kennzahl und pro Produkt sowohl für das laufende Jahr als auch im Vergleich zu Vorjahren benötigt.
- Es sind aussagekräftige und überregional vergleichbare Kennzahlen für die Nutzung des Gesamtangebotes erforderlich.
- Mit Hilfe geeigneter Tools müssen ad-hoc-Fragestellungen sowohl auf Produktebene als auch über das Gesamtangebot bearbeitet werden können.
- Die Analysen müssen regelmäßig, z.B. monatlich, fortgeschrieben werden.
- Die Analysedaten müssen in weiterverwertbarer Form, z.B. zur Einbindung in Berichte, bereitgestellt werden.

2.3.2 Eingrenzung der Variablen und Parameter

Aufgrund der in Kapitel 2.2 beschriebenen Randbedingungen wird empfohlen, die routinemäßige Ermittlung von Kennzahlen auf folgende Variablen und Parameter einzugrenzen:

- Als zentrale Kennzahl der Nutzung wird die Variable
  Anzahl PageViews pro Schlüssel-URL pro Zeiteinheit
  verwendet.
  Schlüssel-URL ist dabei ein einzelnes, vom Client anklickbares Element, welches im Feld cs-uri-stem des WWW-Logs aufgezeichnet wird.
  Schlüssel-URLs werden z.B. von den Fachabteilungen als repräsentative Seiten für die Beurteilung von Nutzungen benannt.

Als Zeiteinheit wird das Monatsraster der Teile "Jahr" und "Monat" des Datum-Feldes des WWW-Logs empfohlen.

- Ergänzende Informationen aus der Log-Datei, wie Abfrage-Parameter (cs-uri-query), Befehlstyp (cs-method) etc. werden ignoriert.

- Es werden nur Datensätze mit den Statuscodes (sc-status) 200, 206 und 304 berücksichtigt.

- Mit Ausnahme einer PageView-Summe über alle Inhalts-Seiten des GESIS-Angebotes wird auf die Ermittlung von Page-Views für Gruppen von Seiten verzichtet, da diese Zahlen nur eine geringe vergleichende Relevanz besitzen.

- Es wird wegen der in Kap. 2.2 geschilderten Probleme in dieser Untersuchung darauf verzichtet, Besucher oder Besuchergruppen zu individualisieren.
  Es ist noch zu prüfen, ob Aussagen zur Anzahl der unterschiedlichen Besucher und deren Gliederung nach Domänen eine ausreichende Relevanz besitzen.

- Für die operative Nutzungs-Analyse werden mit Hilfe des Feldes c-ip aus den Log-Dateien alle Zugriffe
  von Suchmaschinen,
  von IP-Adressen der GESIS-Insitute
  herausgefiltert.
  Zugriffe von Proxy-Servern werden als Indiz für triggernde Benutzeranfragen unverändert berücksichtigt.

D.h.: Aus der WWW-Log-Datei-Struktur (s. Kap. 2.2.2) werden zunächst nur folgende Felder systematisch ausgewertet:

- date
- c-ip
- cs-uri-stem
- sc-status

2.4 Schwächen des Verfahrens "WebSuxess 4"

Eine wesentliche Schwäche des Verfahrens WebSuxess 4 besteht darin, dass zwar innerhalb des Verfahrens eine Vielzahl einschließender und ausschließender Filtermöglichkeiten angeboten werden, dass der zugehörige Input aber nicht selektiv und nicht aus Standard-Datei-Formaten importiert werden kann.

So ist u.a. die manuelle Pflege von Filtern für rund 5.000 Suchmaschinen-IP-Adressen unzumutbar, zumal diese Pflege an 2 Stellen erforderlich wäre: In den Dateien search2.dat und spiders.dat zur Klassifikation der Adressen als Suchmaschinen und im Filter für auszuschließende IP-Adressen.

Ein weiterer Mangel ist die fehlende Möglichkeit, den zeitlichen Verlauf z.B. der PageViews für mehrere einzelne Seiten getrennt aber parallel innerhalb eines Verfahrensablaufs darzustellen. Darstellbar ist in WebSuxess nur der zeitliche Verlauf über die Summe der im laufenden Verfahren aktivierten Seiten. Zwar können in WebSuxess einzelne Seiten per einschließendem Filter separiert ausgewertet werden, hierzu ist aber pro Filtereinstellung jeweils ein neuer Verfahrensdurchlauf erforderlich.

Störend ist das Fehlen von Konfigurationsparametern, mit denen z.B. die Groß-/Kleischreibung der Client-Anfragen standardisiert oder Schreibfehler ausgeblendet werden können.

2.5 Vor- und Nachteile von Filter-Prozeduren

Die Voruntersuchungen haben ergeben, dass es erforderlich ist, vor einer Analyse der Daten mit einem Tool wie WegSuxess mit Hilfe von Filterprozeduren sowohl nicht erwünschte Datensätze herauszufiltern als auch die Datensätze mit den gewünschten Zielinformationen zu selektieren.

Hierzu wurden 5 Prozedurtypen entwickelt (Beschreibung im folgenden Kapitel), die jeweils auf den Gesamt-Datenbestand oder auf Teilmengen angewendet werden können:

- Trennung der Datensätze mit einer c-ip (IP-Adresse des Client) aus der Liste der Suchmaschinen von den übrigen Datensätzen, Erzeugung von 2 Ergebnis-Dateien im IIS-WWW-Log-Format, die separat ausgewertet werden.

- Trennung der Datensätze mit einer c-ip der GESIS-Institute, Erzeugung von 2 Ergebnis-Dateien im IIS-WWW-Log-Format, die separat ausgewertet werden.

- Selektion der Datensätze mit den Ziel-URLs (cs-uri-stem), für die Page-View-Auswertungen erstellt werden sollen, in eine Ergebnis-Datei mit IIS-WWW-Log-Format.

- Umformatierung des IIS-WWW-Log-Formates einer Datei in ein Import-Format mit Komma-getrennten Feldern für die Weiterverarbeitung mit Excel, Datenbanken (z.B. MS Access), SPSS o.ä.

- Erstellung von Zeitreihen mit monatlichem Raster für PageViews auf Ziel-URLs.

Abgesehen von der aktuellen Notwendigkeit der Filterung unerwünschter Datensätze im Hinblick auf die Schwächen des verfügbaren Auswertungs-Tools besteht der Vorteil der Filterung und Selektion in Ergebnis-Dateien vor allem darin, dass die Datenmengen für die weitere Auswertung erheblich keiner sind als der Basis-Datenbestand. Die Weiterverarbeitung wird damit performanter und die Ergebnis-Präsentation übersichtlicher.

Der einzige schwerwiegende Nachteil insbesondere der Selektion von Datensätzen mit definierten Ziel-URLs besteht darin, dass hierbei die zusammenhängende Sicht auf Visits zerstört wird, da Sprünge in und innerhalb nichtselektierter Teile des Bestandes für die Visit-Analyse nicht mehr verfügbar sind. Damit sind dann auch Referrer- und Navigations-Untersuchungen in dem selektierten Datenbestand nur noch begrenzt sinnvoll.

3 Eigenschaften der Filter-Prozeduren

3.1 Vorbemerkung

Das Auswertungs-Tool WebSuxess ist nach überwiegender Auffassung einschlägiger Anwender eines der besten mit vertretbaren Kosten verfügbare Werkzeuge zur Auswertung von WWW-Log-Dateien auf dem Markt. Unter diesem Gesichtspunkt muss mit den Schwächen dieses Systems (oder anderen Schwächen anderer Produkte) gelebt werden.

Die Notwendigkeit zur Erstellung von Filtern vor einer Auswertung der Daten mit WebSuxess, z.B. für die Ausgrenzung der Zugriffe von nicht in der Auswertung erwünschten Client-Adressen (u.a. der regelmäßigen Abfragen des Leitstandes im IZ Bonn), ergab sich bereits mit der Einführung einer systematischen Auswertung der WWW-Log-Dateien.

Es zeigt sich allerdings zunehmend, dass auch ein gutes Auswertungs-Standard-Tool zwar in der Lage ist, sehr flexibel hochkomplexe Auswertungen durchzuführen, dass dafür aber Schwächen einerseits bei der Ab- und Eingrenzung der Ziel-Objekte, andererseits bei der Behandlung regelmäßige Routineaufgaben in Kauf genommen werden müssen.

Da sich mit zunehmender Unschärfe ursprünglich wichtiger Log-Parameter, z.B. der IP-Adresse als Benutzer-Identifikation, der Spielraum für die Ermittlung sinnvoller Kennzahlen einengt, stellt sich die Frage, ob für Routineauswertungen nicht (wieder) spezielle Prozeduren anstelle oder ergänzend zu einem komplexen Auswertungs-Tool eingesetzt werden sollten.

Die Frage verschärft sich, da die auszuwertenden Log-Dateien bei näherer Betrachtung eine sehr schlichte "Flat-File-Struktur" mit nur wenigen relevanten Feldern besitzen und speziell für diese Art von Daten Programmiersprachen wie Perl, AWK o.ä. zur Verfügung stehen, mit denen sowohl differenzierte Auswertungen als auch Umformatierungen in Import-Formate für Auswertungen in Excel, SPSS o.ä. mit geringem Aufwand erstellt werden können.

Die vorliegende Untersuchung enthält mit den im folgenden beschriebenen Prozeduren einen Vorschlag für einen Satz von Verfahrensteilen, aus denen regelmäßige Routineauswertungen zusammengestellt werden können.

Wegen des besonders geringen Programmieraufwandes wurden die Prototypen der folgenden Prozeduren in AWK programmiert und in einer Batch-Umgebung ausgeführt. Für die Überführung in einen Routineeinsatz empfiehlt sich die Recodierung z.B. in Perl.

Anmerkung zur vorliegenden AWK-Programmierung (um gutgemeinten Ratschlägen vorzubeugen):
Das Hauptaugenmerk bei der Erstellung der vorliegenden Programme lag auf der Erzeugung korrekter Ergebnisse. Hinsichtlich Programmierstil und Performance kann sicher noch viel verbessert werden.

3.2 Suchmaschinen

Als Input für den Suchmaschinen-Filter wird eine ASCII-Datei (suchma-ip.txt) verwendet, die pro Record die vollständige IP-Adresse eines (potentiellen) Suchmaschinen-Roboters enthält. Die Adressen werden zu Beginn der Prozedur in ein Array (suma[ ]) eingelesen.

Beim anschließenden sequentiellen Lesen der Input-Log-Datei werden 2 Output-Dateien erzeugt

- exs.log: Log-Records von Suchmaschinen,
- ex1.log: sonstige Log-Records

Die Output-Dateien werden von einer Batch-Prozedur umbenannt in das Auswertungs-Unterverzeichnis kopiert.

Die Selektion erfolgt nach folgenden Regeln:

- Alle von IIS erzeugten Kommentar-Records (1. Zeichen = #) werden zur Erhaltung der Format-Kompatibilität in beide Dateien kopiert.
- Records, deren 3. Feld (c-ip) mit einem Array-Eintrag identisch ist, werden in exs.log, geschrieben, nachdem alle Zeichen auf Kleinschreibung umgesetzt wurden.
- alle sonstigen Records werden in ex1.log geschrieben, nachdem alle Zeichen auf Kleinschreibung umgesetzt wurden.

suchma.awk

```
BEGIN      {while ((getline < "suchma-ip.txt") > 0) {suma[$1]="1"}}
           {if (substr($1,1,1)=="\#") {
               print >> "exs.log"
               print >> "ex1.log"}
            else {
               if (suma[$3]=="1") {$0=tolower($0);print >> "exs.log"}
               else {$0=tolower($0); print >> "ex1.log"}}
           }
```

Die Input-Datei suchma-ip.txt enthält z.Z. 4.590 IP-Adressen von potentiellen Suchmaschinen-Robotern in folgendem Format:

suchma-ip.txt (Ausschnitt)

```
129.250.233.33
142.75.65.141
144.140.254.227
149.174.106.123
149.174.34.13
151.189.12.147
166.90.205.55
171.64.75.83
192.215.220.3
192.41.33.220
192.49.214.115
193.102.192.185
193.14.34.204
193.189.238.2
193.189.238.7
...
216.35.103.1
216.35.103.2
216.35.103.3
216.35.103.4
216.35.103.5
216.35.103.6
216.35.103.7
216.35.103.8
216.35.103.9
216.35.103.10
...
```

Tatsächlich gefunden wurden in der WWW-Log-Datei des Jahres 2001 (s. Kap. 5.2.3.):

- 379 unterschiedliche Roboter
- mit zusammen rund 0,7 Mio PageViews

bei einer Gesamtzahl von (s. Kap. 5.2.1)

- rund 176.000 "Besuchern" und
- zusammen rund 5 Mio PageViews

Eine Detailanalyse der Ergebnisse erfolgt in Kap. 5.

3.3 GESIS-Nutzung

Der GESIS-Filter ist technisch identisch mit dem Suchmaschinen-Filter und verwendet als Input die ASCII-Datei gesis-ip.txt, die alle IP-Adressen der 3 Subnetze

- 193.175.238.* (IZ Bonn)
- 193.175.239.* (GESIS Außenstelle Berlin)
- 193.196.10.* (ZUMA Mannheim)

enthält.

Für das ZA Köln war eine gezielte Filterung nicht möglich, da hier das B-Netz der Uni-Köln (134.95.*.*) mitbenutzt wird. Zwar konnten einige Hosts mit DNS-Namen aus der Domäne za.uni-koeln.de in dem Subnetz 134.95.45.* identifiziert werden, deren Zugriffsvolumen allerdings unbedeutend war. Häufungen von Zugriffen bestimmter, nicht einem offensichtlichen Server-Namen zuzuordnenden Besucher des Netzes 134.95.*.* hatten jeweils ein für die Auswertung unbedeutendes Zugriffsvolumen und wurden deshalb in der GESIS-Filter-Liste nicht berücksichtigt..

Anmerkung:
Rückfragen beim ZA ergaben, dass dort ein DHCP-Server für den Adress-Pool 134.95.45.* eingesetzt wird, dies wird bei zukünftigen Auswertungen berücksichtigt werden.

Die Adressen werden zu Beginn der Prozedur in ein Array (suma[ ]) eingelesen.

suchgesis.awk

```
BEGIN      {while ((getline < "gesis-ip.txt") > 0) {suma[$1]="1"}}
           {if (substr($1,1,1)=="\#") {
               print >> "exs.log"
               print >> "ex1.log"}
            else {
               if (suma[$3]=="1") {$0=tolower($0);print >> "exs.log"}
               else {$0=tolower($0); print >> "ex1.log"}}
           }
```

Beim anschließenden sequentiellen Lesen der Input-Log-Datei werden 2 Output-Dateien erzeugt

- exs.log: Log-Records von GESIS-Zugriffen,
- ex1.log: sonstige Log-Records

Die Output-Dateien werden von einer Batch-Prozedur umbenannt in das Auswertungs-Unterverzeichnis kopiert.

Die Selektion erfolgt nach folgenden Regeln:

- Alle von IIS erzeugten Kommentar-Records (1. Zeichen = #) werden zur Erhaltung der Format-Kompatibilität in beide Dateien kopiert.
- Records, deren 3. Feld (c-ip) mit einem Array-Eintrag identisch ist, werden in exs.log, geschrieben, nachdem alle Zeichen auf Kleinschreibung umgesetzt wurden.
- alle sonstigen Records werden in ex1.log geschrieben, nachdem alle Zeichen auf Kleinschreibung umgesetzt wurden.

Die GESIS-Filter-Liste enthält 765 Einträge potentieller GESIS-Adressen.

Tatsächlich gefunden wurden in der WWW-Log-Datei des Jahres 2001 (s. Kap. 5.2.4):

- 280 unterschiedliche GESIS-Besucher
- mit zusammen rund 1,2 Mio PageViews

Spitzenreiter sind die automatischen, regelmäßigen Zugriffe der Test- und Pflegedienste in Berlin und Bonn mit jeweils bis zu 100-500.000 PageViews.

Eine Detailanalyse der Ergebnisse erfolgt in Kap. 5.

3.4 URL-Selektion

Die URL-Filterung hat das Ziel, alle Zugriffe auf die unter operativen Gesichtspunkten ausgewählten Schlüssel-URLs aus dem Gesamtbestand herauszuselekieren, um sie gezielt und ungestört weiter verarbeiten zu können.

Die Methode ähnelt dem Suchmaschinen-Filter, ist aber etwas komplexer:

Zunächst wird die ASCII-Datei filter-url.txt eingelesen, die pro Record eine der Schlüssel-URLs enthält, diese Zeichenketten werden in ein Arry (filter[ ]) kopiert. Zusätzlich wird die ASCII-Datei filter-out eingelesen, die pro Record eine Zeichenkette enthält, deren auftreten in einer URL diese als nicht-auszuwertend kennzeichnet. Diese Records werden in den Array filterout[ ] kopiert.

Als prinzipielles Selektionskriterium werden diejenigen Statuscodes (Feld 8, sc-status) in den Variablen code1, code2 ff festgelegt, die als Voraussetzung für eine Selektion definiert werden.

Als Daten-Eingabe wird üblicherweise die "Sonstige"-Datei als Ergebnis der Suchmaschinen+GESIS-Filter-Prozeduren (filb2001all.log, s. Kap. 4) verwendet.

Als Ergebnis werden 3 Output-Dateien erzeugt:

- filb.log mit den selektierten Zugriffen auf die Schlüssel-URLs,
- fil-code.log mit den Log-Records, die einen falschen Statuscode enthalten,
- fil-out.log mit den Log-Records, die wegen einer filter-out-Zeichenkette in Feld 6 nicht selektiert wurden.

Die Output-Dateien werden von einer Batch-Prozedur umbenannt in das Auswertungs-Unterverzeichnis kopiert.

filter.awk

```
BEGIN      {while ((getline < "filter-url.txt") > 0) {filter[$0]=$0}
           while ((getline < "filter-out.txt") > 0) {filterout[$0]=$0}
           code1 = "200"
           code2 = "206"
           code3 = "304"
           }
           {inrec=inrec+1
            # print (inrec "   " outrec  "    " $8)
            if (substr($1,1,1)=="\#") {
              print >> "filb.log"
              }
            else
              {if (($8!~code1)&&($8!~code2)&&($8!~code3)) {
                 print >> "fil-code.log"
                 next
                 }
               $7 = "-"
               $11= "-"
               $0=tolower($0)
               for (i in filter) {
                 zzz=filter[i]
                 if ($0 ~ zzz) {
                    for (j in filterout) {
                       out=filterout[j]
                       if ($6 ~ out) {
                          print >> "fil-out.log"
                          next
                          }
                    }
                    print >> "filb.log"
                    outrec=outrec+1
                    next
                    }
                 }
              }
           }
```

Die Selektion erfolgt nach folgenden Regeln:

- Alle von IIS erzeugten Kommentar-Records (1. Zeichen = #) werden zur Erhaltung der Format-Kompatibilität in die Output-Datei filb.log kopiert.

- Es wird geprüft, ob Feld 8 (sc-status) einen der gewünschten Status-Codes enthält. Wenn ja, wird die Selektion fortgesetzt, wenn nein, wird der nächste Record eingelesen.

- Der bearbeitete Input-Record ($0) wird wie folgt standardisiert:
  - Das Feld 7 (cs-uri-query) wird auf den Standardwert "-" gesetzt,
  - das Feld 11 (cs(Referes) wird auf den Standardwert "-" gesetzt,
  - alle Zeichen werden auf Kleinschreibung standardisiert,

- Für alle Elemente des filter-Array wird geprüft, ob deren Inhalt als Zeichenkette in dem standardisierten Input-Record vorkommt. Wird eine Ü-

bereinstimmung gefunden, so wird die Selektion fortgesetzt, wenn nein, wird der nächste Record eingelesen.

- Für alle Elemente des filterout-Array wird geprüft, ob diese Zeichenkette in Feld 6 des Records vorkommt. bei negativem Ergebnis gilt der Record als selektiert und wird in die Output-Datei filb.log geschrieben.

Eine Besonderheit der Filter-Prozedur gegenüber der bei dem Suchmaschinen-Filter verwendeten Methode besteht darin, dass der Zeichenketten-Vergleich mit den filter-Array-Inhalten nicht nur für das Ziel-Feld 6 des Input-Records sondern für den gesamten Record - unter Einschluss der Leerzeichen vor und nach Feld 6 - durchgeführt wird.

Dies ist eine Methode, um als Schlüssel-URL nicht nur explizite Dateinamen (z.B. `/information/foris/index.htm`) sondern auch Teilketten, z.B. Unterverzeichnisse einschließlich deren Untenmengen (z.B. `/information/themen/fokus/`)

zu definieren:

Die Schreibweise der Schlüssel-URLs in der Datei filter-url.txt folgt folgenden Regeln:

- In der Input-Datei filter-url.txt beginnen alle URL-Zeichenketten mit einem Leerzeichen, um sicherzustellen, das diese Zeichenkette mit Feld 6, beginnend mit dessen 1. Zeichen nach einem Leerzeichen verglichen wird.

- Die Kennzeichnung eines expliziten Dateinamens als Schlüssel-URL erfolgt in der Datei filter-url.txt durch das Abschließen der Zeichenkette mit einem Leerzeichen, um sicherzustellen, dass diese Zeichenkette auf Identität mit dem gesamten Feld 6 gefolgt von einem Leerzeichen geprüft wird.

- Ein Unterverzeichnis wird in filter-url.txt dadurch gekennzeichnet, dass das abschließende Leerzeichen entfällt, damit wird diese Zeichenkette als Substring des Feldes 6 (beginnend mit dessen 1. Zeichen) getestet.

Anmerkung:

Die Selektion von Unterverzeichnissen einschließlich aller enthaltenen Seiten hat sich nach ersten Produktionsläufen als unzweckmäßig erwiesen, da zu viele unerwünschte Detail-Informations-Seiten berücksichtigt werden. Die aktuelle Filter-Liste (s.u.) verzichtet deshalb auf dieses Merkmal.

Es ist statt dessen vorgesehen, in der nächsten Version des Verfahrens den URL-Filter um eine "Mittel-Trunkierung" zu ergänzen, die eine URL-Gruppe anhand eines Anfangs- und eines End-Strings identifiziert. Auf diese Weise sollen dann neue WWW-Angebote, z.B. im Unterverzeichnis /information/themen/fokus/... erkannt, hierbei aber z.B. nur Seiten mit bestimmten Endungen, z.B. "index.htm", ".doc", ".pdf" berücksichtigt werden.

filter-url.txt (Stand 30.1.2001)

```
 /bestellen/iz/index.htm
 /en/cooperation/information/eastern_europe/index.htm
 /en/cooperation/information/eastern_europe/associ.shtm
 /en/cooperation/information/eastern_europe/network_ee.htm
 /en/events/transformation/index.htm
 /en/gesis_branch_office/index.htm
 /en/gesis_branch_office/overview.htm
 /en/information/eastern_europe/ineaste/index.htm
 /en/information/eastern_europe/proeaste/index.htm
 /en/information/eastern_europe/proeaste/search/index.htm
 /en/information/eastern_europe/proeaste/survey/index.htm
 /en/information/journals/eastern_europe/index.htm
 /en/information/theme/europa.htm
 /en/publications/magazines/newsletter_eastern_europe/index.htm
 /en/publications/magazines/newsletter_eastern_europe/nl014/index.htm
 /en/publications/magazines/newsletter_eastern_europe/nl014/nl014.pdf
 /en/publications/magazines/newsletter_eastern_europe/nl013/index.htm
 /en/publications/magazines/newsletter_eastern_europe/nl013/nl013.pdf
 /en/publications/magazines/newsletter_eastern_europe/nl01s/index.htm

/en/publications/magazines/newsletter_eastern_europe/nl01s/nl_sh_2001.pdf
 /en/publications/magazines/newsletter_eastern_europe/nl012/index.htm
 /en/publications/magazines/newsletter_eastern_europe/nl012/nl012.pdf
 /en/publications/magazines/newsletter_eastern_europe/nl011/index.htm
 /en/publications/magazines/newsletter_eastern_europe/nl011/nl011.pdf
 /en/publications/magazines/newsletter_eastern_europe/nl004/index.htm
 /en/publications/magazines/newsletter_eastern_europe/nl004/nl004.pdf
 /en/publications/magazines/newsletter_eastern_europe/nl003/index.htm
 /en/publications/magazines/newsletter_eastern_europe/nl003/nl003.pdf
 /en/publications/magazines/newsletter_eastern_europe/nl002/index.htm
 /en/publications/magazines/newsletter_eastern_europe/nl002/nl002.pdf
 /en/publications/magazines/newsletter_eastern_europe/nl001/index.htm
 /en/publications/magazines/newsletter_eastern_europe/nl001/nl001.pdf
 /en/staff/as/index.htm
 /gesis_aussenstelle/index.htm
 /gesis_aussenstelle/uebersicht.htm
 /index.htm
 /iz/index.htm
 /information/index.htm
 /information/foris/erhebung/index.htm
 /information/foris/erhebung/preview/index.htm
 /information/foris/index.htm
 /information/foris/recherche/index.htm
 /information/osteuropa/ineaste/index.htm
 /information/osteuropa/proeaste/erhebung/index.shtm
 /information/osteuropa/proeaste/index.htm
 /information/osteuropa/proeaste/recherche/index.htm
 /information/rechercheunterst/klassifikation/index.htm
 /information/rechercheunterst/klassifikation/klass.htm
 /information/rechercheunterst/klassifikation/klass.pdf
```

```
/information/sofo/index.htm
/information/sofo/recherche/index.htm
/information/solis/index.htm
/information/themen/europa.htm
/information/themen/fokus/balkan/index.htm
/information/themen/fokus/balkan/balkan.pdf
/information/themen/fokus/bse/index.htm
/information/themen/fokus/bse/bse.pdf
/information/themen/fokus/eu/index.htm
/information/themen/fokus/eu/eu.pdf
/information/themen/fokus/euro/index.htm
/information/themen/fokus/euro/euro.pdf
/information/themen/fokus/gesundheit/index.htm
/information/themen/fokus/gesundheit/gesundheit.pdf
/information/themen/fokus/h_gew/index.htm
/information/themen/fokus/h_gew/h_gew.pdf
/information/themen/fokus/islam/index.htm
/information/themen/fokus/islam/islam.pdf
/information/themen/fokus/klima/index.htm
/information/themen/fokus/klima/klima.pdf
/information/themen/fokus/pol_skan/index.htm
/information/themen/fokus/pol_skan/pol_skan.pdf
/information/themen/fokus/rechtsradikalismus/index.htm
/information/themen/fokus/rechtsradikalismus/rechtsradikalismus.pdf
/information/themen/fokus/renten/index.htm
/information/themen/fokus/renten/renten.pdf
/information/themen/fokus/school/index.htm
/information/themen/fokus/school/school.pdf
/information/themen/fokus/solda/index.htm
/information/themen/fokus/solda/solda.pdf
/information/themen/fokusplus/index.htm
/information/themen/fokusplus/amerika/index.htm
/information/zeitschriften/deutschspr/index.htm
/information/zeitschriften/deutschspr/zeitschriften.doc
/information/zeitschriften/deutschspr/zeitschriften.pdf
/information/zeitschriften/deutschspr/zsliste_gesamt.htm
/information/zeitschriften/index.htm
/information/zeitschriften/osteuropa/index.shtm
/kooperation/information/osteuropa/associ.htm
/kooperation/information/osteuropa/index.htm
/kooperation/information/osteuropa/presse.htm
/mitarbeiter/as/index.htm
/veranstaltungen/transformationsforschung/index.htm
```

Die in dieser Untersuchung verwendete Filterliste enthält 92 URL-Regeln.

Für diesen inhaltlichen Ausschnitt gefunden wurden (s. Kap. 5.3.1.):

- 92 unterschiedliche Seiten
- 262.000 PageViews
- 63.500 "Besucher" (ohne Suchmaschinen, ohne GESIS)

Die filter-out-Datei enthält Zeichenketten, die - bei der Auflösung der Unterverzeichnisse aus der filter-url-Liste auftretende - nicht-erwünschte URLs identifizieren und ausfiltern sollen.

filter-out.txt

```
_gesis_tools
_vti_bin
.js
noframe.htm
.ico
.gif
http:
www.gesis.org
```

Die filter-out-Liste identifiziert

- bestimmte Datei-Typen (.js, .ico, .gif),
- Navigations-Hilfsmittel (_gesis_tools etc.),
- Schreibfehler des Client, die vom Server nicht als solche erkannt wurden.

Der Statuscode-Filter schreibt Records mit nicht-gültigem Statuscode in die Datei fil-code.log.

Da bei einer Größe der Gesamt-WWW-Log-Datei von rund 2,3 GB die Datei fil-code.log eine Größe von knapp 0,2 GB hat, kann auf eine Rate von rund 8% fehlerhafter Zugriffe auf den Server geschlossen werden.

3.5 Import-Format für Excel o.ä.

Als Import-Format u.a. für Excel biete sich das csv-Format (coma separated values) an, ein ASCII-Format, bei dem die einzelnen Felder eines Records durch Komma getrennt werden.

Es wird eine Prozedur bereitgestellt, mit der beliebige WWW-Log-Format-lompatible Dateien in folgender Weise umformatiert werden können:

makecsv.awk

```
        {if (substr($1,1,1)!="\#") {
             monat=substr($1,1,7)
             ip=$3
             url=$6
             print (monat "," ip "," url) >> "exx.log"
             }
        }
```

- Die mit einem "#" beginnenden Sätze werden ignoriert,
- Output-Feld 1 ist der Substring Jahr-Monat des Input-Date-Feldes,
- Output-Feld 2 ist das Feld 3 des Input-Satzes mit der IP-Adresse des Client,
- Output-Feld 3 ist das Feld 6 des Input-Satzes mit der URL.

Zweck der Prozedur ist es vorrangig, Nachuntersuchungen, Sortierungen, Pivot-Analysen etc. an Selektionsergebnissen u.a. mit Excel durchzuführen,

3.6 Zeitreihen

Mit der Zeitreihen-Prozedur soll eine Übersichts-Ausgabe erstellt werden, bei der pro selektierter URL je eine Zeile mit den Feldern

- URL
- Summe der PageViews
- PageViews pro Monat für alle Monate des Auswertungszeitraumes

erzeugt wird.

zeitreihe.awk

```
{if (substr($1,1,1)!="\#") {
              monat=substr($1,1,7)
              ip=$3
              url=$6
              zr[url,monat] = zr[url,monat]+1
              mo[monat]=monat
              ur[url]=url
              inrec=inrec+1
              # print (inrec, monat, zr[url,monat])
              }
            }
END         {anzur = asort(ur,urso)
             anzmo = asort(mo,moso)
             for  (i=1;i<=anzmo;i++)  {print  (i  "     "  moso[i])  >>  "zeit-
moso.csv"}
             for  (i=1;i<=anzur;i++)  {print  (i  "     "  urso[i])  >>  "zeit-
urso.csv"}
             monum=""
             for(j=1;j<=anzmo;j++) {
                monum=(monum "," moso[j])
                }
             print ("URL" ",Summe" monum) >> "zeitreihe.csv"
             for (i=1;i<=anzur;i++) {
                monall=""
                jahrsum=0
                for (j=1;j<=anzmo;j++) {
                   monall=(monall  ","  zr[urso[i],moso[j]])
                   jahrsum=jahrsum+zr[urso[i],moso[j]]
                   }
                print (urso[i] "," jahrsum  monall) >> "zeitreihe.csv"
                }
            }
```

Input-Datei ist üblicherweise die Ergebnis-Datei filb.log der URL-Selektion.

Es werden folgende Felder aus der Input-Datei selektiert:

- Die mit einem "#" beginnenden Sätze werden ignoriert,
- monat = Substring Jahr-Monat des Input-Date-Feldes,
- url = Feld 6 des Input-Satzes mit der URL.

Pro eingelesenem Input-Record wird das AWK-Array-Element zr[url,monat] um 1 hochgezählt.

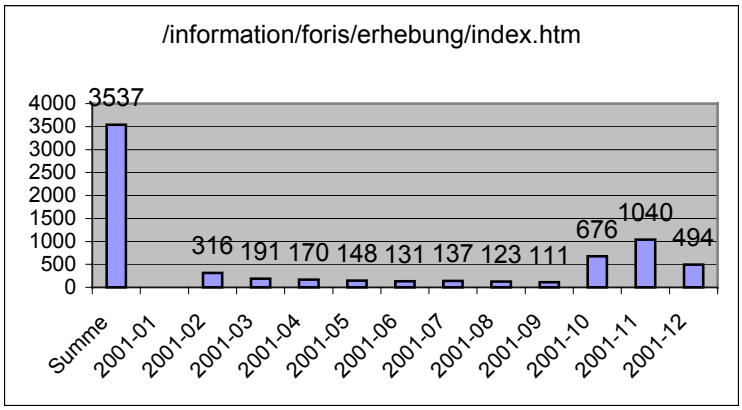
Abschließend werden die im Array gesammelten Ergebnisse sortiert und mit einer URL und deren monatlichen PageViews-Zahlen pro Zeile csv-formatiert in die Datei zeitreihe.csv ausgegeben.

Die Datei kann unmittelbar mit Excel aufgerufen und angezeigt werden.

zeitreihe.csv, Excel-Anzeige Ausschnitt

URL	Summe	2001-01	2001-02	2001-03	2001-04	2001-05	2001-06
/bestellen/iz/index.htm	22678		1035	705	2349	503	461
/information/foris/erhebung/index.htm	3537		316	191	170	148	131
/information/foris/erhebung/preview/index.htm	766		54	65	58	43	29
/information/foris/index.htm	5139	1	523	407	432	445	366
/information/foris/recherche/index.htm	8081		764	695	640	633	555
/information/rechercheunterst/klassifikation/10000.htm	7				1	1	2
/information/rechercheunterst/klassifikation/102.htm	482		51	37	57	34	25
/information/rechercheunterst/klassifikation/10200.htm	41				2	3	7
/information/rechercheunterst/klassifikation/103.htm	104		11	7	15	10	8
/information/rechercheunterst/klassifikation/10300.htm	14				1	2	2
/information/rechercheunterst/klassifikation/104.htm	58		9	4	9	3	3
/information/rechercheunterst/klassifikation/10400.htm	1						
/information/rechercheunterst/klassifikation/105.htm	164		11	18	17	18	14
/information/rechercheunterst/klassifikation/10500.htm	53				1	7	6

In Excel ist es bei Bedarf möglich, mit wenigen Handgriffen die Monats-Verläufe ausgewählter Zeilen graphisch darzustellen:



4 Verfahrens-Empfehlungen

4.1 Vorbemerkung

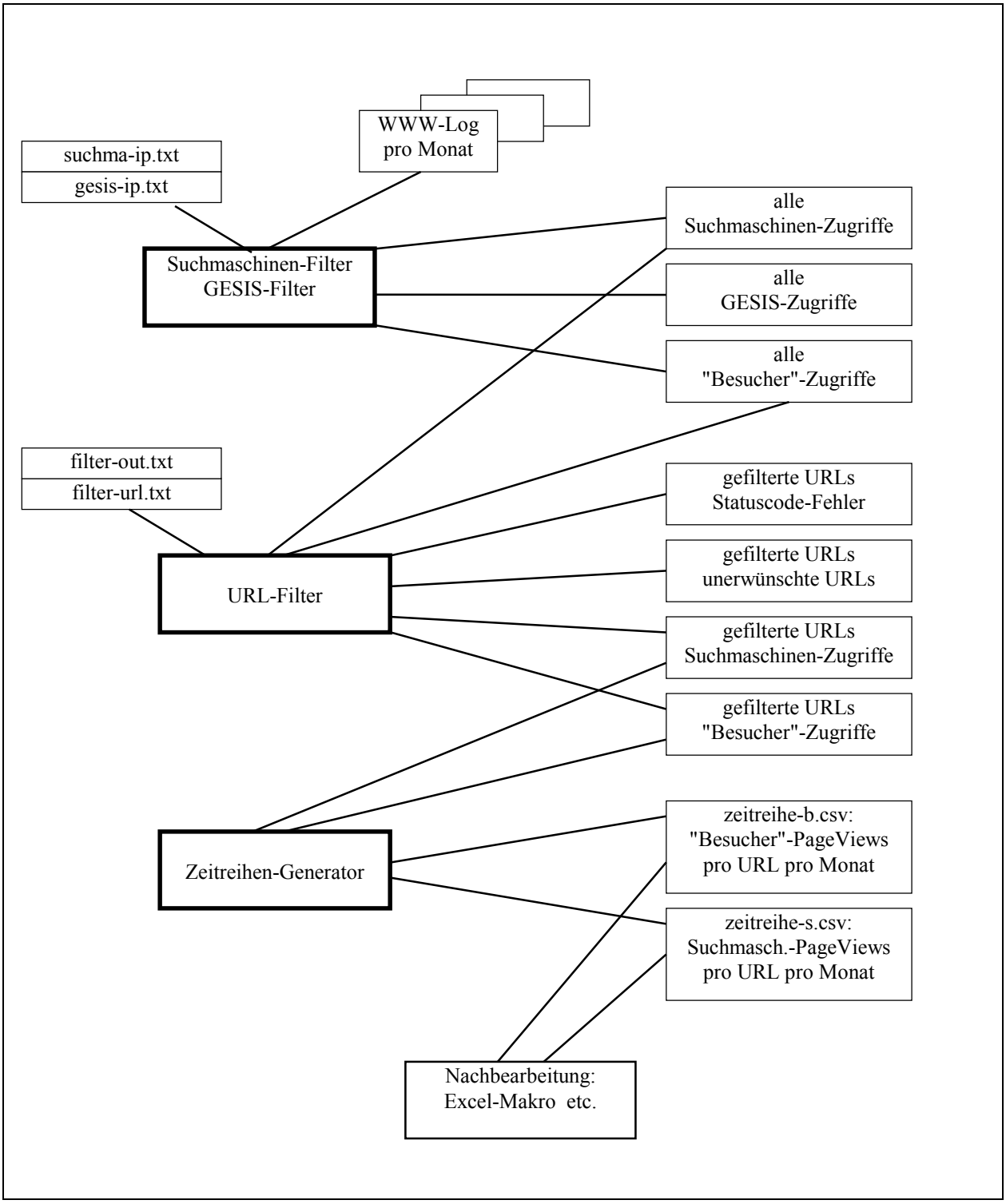
Mit Hilfe der von den Filter-Prozeduren erzeugten Selektionen stehen insgesamt folgende Dateien für Auswertungen zur Verfügung (die u.g. Dateinamen sind die in dieser Auswertung verwendeten Bezeichnungen nach dem Umspeichern der Ergebnisse der AWK-Prozeduren in das Auswertungsverzeichnis) (die Kennung (F) bezeichnet Dateien, die Format-kompatibel mit der Ursprungsdatei und damit auswertbar mit WebSuxess sind):

- iis2001.log Gesamt-WWW-Log (F)
- fils2001all.log alle Suchmaschinen-Zugriffe (F)
- filz2001all.log Gesamt-WWW-Log ohne Suchmaschinen-Zugriffe (F)
- filg2001all.log alle GESIS-Zugriffe (F)
- filb2001all.log alle "Besucher"-Zugriffe (F)
  (ohne Suchmaschinen, ohne GESIS)
- filb2001-code.log alle Zugriffe mit Statuscode-Fehler
- fila2001.log alle Zugriffe auf gefilterte URLs (F)
  (inkl. Suchmaschinen und GESIS)
- filb2001.log "Besucher"-Zugriffe auf gefilterte URLs (F)
- fils2001.log Suchmaschinen-Zugriffe auf gefilterte URLs (F)
- filg2001.log GESIS-Zugriffe auf gefilterte URLs (F)
- zeitreihe-b.csv Zeitreihen-Auswertung von filb.log
- zeitreihe-s.csv Zeitreihen-Auswertung von fils.log
- makecsv.csv CSV-Formatierung einer der o.g. log-Dateien

4.2 Direktauswertung der Log-Dateien mit Filter-Prozeduren

Unter dem Gesichtspunkt der Gewinnung operativer Kennzahlen der WWW-Nutzung wird empfohlen, die Original-WWW-Log-Dateien regelmäßig, zweckmäßig Scheduler-gesteuert in monatlichen Segmenten, in ein IZ-zentral zugängliches Jahrgangs-Verzeichnis zu kopieren und von dort aus weiterzubearbeiten.

Es wird empfohlen, auf jeweils alle Log-Daten eines Jahrgangs - im aktuellen Jahr vom Jahresanfang bis zum letzten abgeschlossenen Monat - die Filter-Prozeduren in folgender Reihenfolge anzuwenden:



Dieser Ablauf wird in der vorliegenden Untersuchung mit folgender Batch-Prozedur realisiert:

```
@echo off
echo Suchmaschinen aus iis nach nach fils-all, fliz-all
del exs.log
del ex1.log
awk -fsuchma.awk c:\daten\gesis-www\gesislog\iis2001\iis2001.log
copy exs.log c:\daten\gesis-www\gefiltert\fils2001all.log
copy ex1.log c:\daten\gesis-www\gefiltert\filz2001all.log
echo GESIS aus filz-all nach filg-all, filb-all
del exs.log
del ex1.log
awk -fsuchgesis.awk c:\daten\gesis-www\gefiltert\filz2001all.log
copy exs.log c:\daten\gesis-www\gefiltert\filg2001all.log
copy ex1.log c:\daten\gesis-www\gefiltert\filb2001all.log
echo URLs aus filb-all nach filb
del filb.log
del fil-code.log
del fil-out.log
awk -ffilter.awk c:\daten\gesis-www\gefiltert\filb2001all.log
copy filb.log     c:\daten\gesis-www\gefiltert\filb2001.log
copy fil-code.log c:\daten\gesis-www\gefiltert\filb2001-code.log
copy fil-out.log  c:\daten\gesis-www\gefiltert\filb2001-out.log
echo URLs aus fils-all nach fils
del filb.log
del fil-code.log
del fil-out.log
awk -ffilter.awk c:\daten\gesis-www\gefiltert\fils2001all.log
copy filb.log     c:\daten\gesis-www\gefiltert\fils2001.log
copy fil-code.log c:\daten\gesis-www\gefiltert\fils2001-code.log
copy fil-out.log  c:\daten\gesis-www\gefiltert\fils2001-out.log
echo zeitreihe Besucher
del zeitreihe.csv
del zeit-moso.csv
del zeit-urso.csv
awk -fzeitreihe.awk c:\daten\gesis-www\gefiltert\filb2001.log
copy zeitreihe.csv c:\daten\gesis-www\gefiltert\zeitreihe-b.csv
echo zeitreihe Suchmaschinen
del zeitreihe.csv
del zeit-moso.csv
del zeit-urso.csv
awk -fzeitreihe.awk c:\daten\gesis-www\gefiltert\fils2001.log
copy zeitreihe.csv c:\daten\gesis-www\gefiltert\zeitreihe-s.csv
```

Es wird empfohlen, die regelmäßigen Routine-Informationen in der Kurzfassung "PageViews/URL/Monat" mit Hilfe der Filter-Prozeduren und des Zeitreihen-Generators direkt aus den Log-Dateien zu erzeugen und den zuständigen Mitarbeiter zugänglich zu machen.

Zu diesem Zweck muss die Filter-Prozedur filter-url.txt (mit einem geeigneter Editor - z.B. Word - zur Berücksichtigung der signifikanten Leerstellen) systematisch gepflegt werden.

Abschätzung des Aufwandes:

Laufzeiten der AWK-Prozeduren auf einem PC-Arbeitsplatz mit 600 MHz-Prozessor über die WWW-Log-Datei des Jahres 2001 mit 2,3 GB:

- Suchmaschinen-Filter plus GESIS-Filter: ca. 60 Minuten
- URL-Filter (pro Durchlauf): ca. 180 Minuten
- Zeitreihen-Generator: ca. 15 Sekunden
- Gesamt-Durchlauf der Batch-Prozedur: ca. 7 Stunden

4.3 Auswertung mit WebSuxess

4.3.1 Auswertung gefilterter Dateien

Die o.g. Filter-Dateien stehen für ergänzende Detail-Analysen zur Verfügung. Es bestehen aus fachlicher Sicht keine Einwände, hierfür WebSuxess zu verwenden.

Es wird empfohlen, bei Bedarf nach individuellen Auswertungen, WebSuxess auf den Arbeitsplätzen der zuständigen Mitarbeiter zu installieren und die Profile für die Auswertung der relevanten, auf einem zentralen Server bereitgestellten Log-Dateien vorzukonfigurieren. Hierbei ist u.a. darauf zu achten, dass die WebSuxess-Parameter - z.B. der Statuscode-Filter und die URL des WWW-Servers - den Anforderungen an die Auswertung entsprechend eingestellt werden.

Nach bisheriger Einschätzung sind folgende Dateien für Detailanalysen interessant:

- alle "Besucher"-Zugriffe (ohne Suchmaschinen und GESIS)
  (u.a. für Visit- und Referrer-Untersuchungen)
- alle Suchmaschinen-Zugriffe
  (u.a. zur Untersuchung der Suchmaschinen-Abdeckung des Angebotes sowie der Auswahl- und Visit-Strategien unterschiedlicher Suchmaschinen)
- gefilterte URLs "Besucher"-Zugriffe
  (u.a. zur Untersuchung der nutzenden "Benutzern" und Domänen)

Auf die in Kap. 2 dargestellten Interpretations-Probleme der Kategorien "Besucher", "Mehrfachbesucher" und "Visit" wird hingewiesen.

4.3.2 Auswertung der Gesamtdatei

Es hat sich gezeigt, dass eine WebSuxess-Analyse der unveränderten WWW-Log-Dateien wegen der erheblichen Einflüsse der Suchmaschinen- und der GESIS-Zugriffe nur bei prinzipiellen Fragestellungen sinnvoll ist.

Als inhaltlich relevante Kennzahlen für die Nutzung des GESIS-WWW-Angebotes werden folgende 2 Werte empfohlen:

- Summe aller erfolgreichen Besucher-PageViews pro Monat / Jahr
  - ohne Suchmaschinen,
  - ohne GESIS/Pflegeaufwand,
  - ohne Graphiken/Navigations-Elemente,
  - mit Statuscode-Filter

  (Datei: alle "Besucher"-Zugriffe, filb2001all.log)
- Summe aller Suchmaschinen-Zugriffe pro Monat / Jahr
  (Datei: alle Suchmaschinen-Zugriffe, fils2001all.log)

Für seriöse Veröffentlichungen sollten beide Zahlen genannt werden, durchaus üblich ist aber auch die Summe beider Zahlen.

Unseriös ist nach Auffassung des Autors

- eine Summe inkl. Pflegeaufwand,
- eine Summe inkl. Navigations-Elemente.

Anmerkung:

Es wird darauf hingewiesen, dass mit zunehmender technischer Raffinesse der Web-Präsentationen der Anteil der Navigationselemente an den Hits erheblich zunimmt.

In den Produktions-Auswertungen 2002 wurden Hits auf folgende Dateitypen als nicht-relevante Navigationselemente ausgefiltert:

- .gif
- .css
- .js
- .dll
- .jpg
- .ico

Die Ergebnisdateien für den Zeitraum 1.1.-13.6.02 hatten folgende Größen (proportional zu der Anzahl der Records):

- Gesamtdatei: 1.586 MB
- Navigationselemente: 1.214 MB (77 % von Gesamt)
- Rest: 372 MB (23 % von Gesamt)
  davon
  Suchmaschinen: 72 MB (20 % von Rest)
  GESIS: 46 MB (12 % von Rest)
  Besucher: 254 MB (68 % von Rest)

5 Exemplarische Ergebnisse

5.1 Vorbemerkung

In den folgenden Abschnitten werden Ergebnisse der Auswertung exemplarisch zusammengestellt und beachtenswerte Zusammenhänge in Kap. 5.6. tabellarisch kommentiert.

5.2 Summenzahlen für WWW.GESIS.ORG

Für die WWW-Log-Datei

- Zeitraumes: 31.01.2001 - 31.12.2001 = 335 Tage
- Größe: rund 2,3 GB

ergeben sich bei einer Darstellung der Ergebnisdateien mit WebSuxess folgende Statistiken:

5.2.1 Gesamt inkl. Suchmaschinen und GESIS (Datei: iis2001.log)

- Gesamtzahl der abgerufenen unterschiedlichen Seiten: 58.438
- Anzahl aller Hits: 13.544.183
- Anzahl PageViews mit Statuscode 200, 206, 304: 5.051.247
- Anzahl unterschiedlicher "Besucher": 176.043
- davon "einmalige Besucher": 142.742 (81%)
- Anzahl Visits: 630.903

5.2.2 "Besucher" (ohne Suchmaschinen, GESIS) (Datei: filb2001all.log)

- Gesamtzahl der abgerufenen unterschiedlichen Seiten: 25.749
- Anzahl aller Hits: 10.737.636
- Anzahl PageViews mit Statuscode 200, 206, 304: 3.097.879
- Anzahl unterschiedlicher "Besucher": 175.386
- davon "einmalige Besucher": 142.700 (81%)
- Anzahl Visits: 466.408

5.2.3 Nur Suchmaschinen (Datei: fils2001all.log)

- Gesamtzahl der abgerufenen unterschiedlichen Seiten: 6.817
- Anzahl aller Hits: 750.870
- Anzahl PageViews mit Statuscode 200, 206, 304: 707.784
- Anzahl unterschiedlicher "Besucher": 379
- davon "einmalige Besucher": 11 (2%)
- Anzahl Visits: 126.163

5.2.4 Nur GESIS (Datei: filg2001all.log)

- Gesamtzahl der abgerufenen unterschiedlichen Seiten: 14.550
- Anzahl aller Hits: 2.055.627
- Anzahl PageViews mit Statuscode 200, 206, 304: 1.245.551
- Anzahl unterschiedlicher "Besucher": 280
- davon "einmalige Besucher": 29 (10%)
- Anzahl Visits: 38.332

5.2.5 Kontroll-Rechnung

Die vorstehenden Ergebnisse wurden durch Präsentation der Ergebnisdateien mit WebSuxess erzeugt. Da es sich bei den Dateien um vollständige und exklusive Teilmengen handelt müssen die Summen der Zahlenwerte für die Parameter

Hits und PageViews,

über die Auswertungen

"aller Besucher" (filb) + "Suchmaschinen" (fils) + "GESIS" (filg)

jeweils mit den entsprechenden Zahlen der Gesamtauswertung (iis) übereinstimmen

Kenngröße	IIS	summme filb+fils+filg	delta iis-summe
Hits	13544183	13544133	50
PageViews	5051247	5051214	33

Wie der Vergleich zeigt, ist dies nicht der Fall, es gibt sowohl bei den Hits als auch bei den PageViews eine - wenn auch sehr geringe - Abweichung. Ursache dieser Unterschiede ist vermutlich die automatische Zeitumstellung Sommerzeit-Winterzeit auf dem Server, die dazu führte, dass Log-Datensätze vorübergehend nicht in der für eine Auswertung mit WebSuxess vorgeschriebenen strikten zeitlichen Reihenfolge der Zeitstempel in den Log-Datensätzen abgespeichert wurden. Beim Einlesen der Dateien in WebSuxess werden dann Datensätze, die nicht dieser Anforderung entsprechen, übersprungen. Die Wirkung dieses Effektes ist in den Teil-Dateien offensichtlich etwas stärken, als in der Gesamt-Datei. Die Differenzen sind allerdings so gering, dass sie für die weitere Auswertung vernachlässigt werden können.

Um diese Fehlerquelle zu vermeiden wäre es zweckmäßig, den Server auf eine konstante Zeit, z.B. UTC, einzustellen. Es ist allerdings unklar, ob hierdurch nicht neue Probleme, z.B. beim Update des WWW-Angebotes mit Frontpage entstehen.

5.3 Schlüssel-URLs

5.3.1 "Besucher" auf Schlüssel-URLs (filb2001.log)

Die folgenden Summenzahlen dienen der Abschätzung, in welchem Umfang die ausgewählten URLs das Nutzungsinteresse der "Besucher" im Vergleich zur Gesamtnutzung repräsentieren.

- Gesamtzahl der abgerufenen unterschiedlichen Seiten: 92
- Anzahl aller Hits: 262.119
- Anzahl PageViews mit Statuscode 200, 206, 304: 262.119
- Anzahl unterschiedlicher "Besucher": 63.529
- davon "einmalige Besucher": 52.498 (82%)
- Anzahl Visits: 124.378

5.3.2 Suchmaschinen Schlüssel-URLs (fils2001.log)

Die folgenden Summenzahlen dienen der Abschätzung, in welchem Umfang speziell die ausgewählten URLs von Suchmaschinen registriert wurden.

- Gesamtzahl der abgerufenen unterschiedlichen Seiten: 91
- Anzahl aller Hits: 26.430
- Anzahl PageViews mit Statuscode 200, 206, 304: 26.430
- Anzahl unterschiedlicher "Besucher": 331
- davon "einmalige Besucher": 22 (6%)
- Anzahl Visits: 13.890

5.3.3 "Besucher"-PageViews von Schlüssel-URLs

Die folgende Tabelle zeigt die Ergebnisse der "Besucher"-PageView-Auszählungen für die ausgewählten URLs als Ausschnitt aus der Zeitreihen-Analyse der Datei filb2001.log.

Die Tabelle zeigt folgende Spalten:

- URL,
- Summe der PageViews im Untersuchungszeitraum,
- PageViews pro Monat für die Monate Oktober, November und Dezember.

URL	Summe	2001-10	2001-11	2001-12
/bestellen/iz/index.htm	22678	492	477	331
/en/cooperation/information/eastern_europe/associ.shtm	79	17	12	14
/en/cooperation/information/eastern_europe/index.htm	143	30	33	20
/en/cooperation/information/eastern_europe/network_ee.htm	35	6	10	7
/en/events/transformation/index.htm	123	33	39	25
/en/gesis_branch_office/index.htm	1718	144	108	97
/en/gesis_branch_office/overview.htm	44	9	12	10
/en/information/eastern_europe/ineaste/index.htm	201	57	58	38
/en/information/eastern_europe/proeaste/index.htm	157	39	46	35
/en/information/eastern_europe/proeaste/search/index.htm	53	14	15	7
/en/information/eastern_europe/proeaste/survey/index.htm	59	18	16	10
/en/information/journals/eastern_europe/index.htm	325	54	47	36
/en/information/theme/europa.htm	121	34	40	24
/en/publications/magazines/newsletter_eastern_europe/index.htm	2587	319	294	127
/en/publications/magazines/newsletter_eastern_europe/nl001/index.htm	455	15	27	12
/en/publications/magazines/newsletter_eastern_europe/nl001/nl001.pdf	513	53	51	34
/en/publications/magazines/newsletter_eastern_europe/nl002/index.htm	346	30	34	25
/en/publications/magazines/newsletter_eastern_europe/nl002/nl002.pdf	1329	142	67	67
/en/publications/magazines/newsletter_eastern_europe/nl003/index.htm	519	20	24	13
/en/publications/magazines/newsletter_eastern_europe/nl003/nl003.pdf	1214	209	71	57
/en/publications/magazines/newsletter_eastern_europe/nl004/index.htm	380	23	28	15
/en/publications/magazines/newsletter_eastern_europe/nl004/nl004.pdf	661	92	61	47

/en/publications/magazines/newsletter_eastern_europe/nl011/index.htm	420	27	30	16
/en/publications/magazines/newsletter_eastern_europe/nl011/nl011.pdf	566	31	41	30
/en/publications/magazines/newsletter_eastern_europe/nl012/index.htm	346	35	32	18
/en/publications/magazines/newsletter_eastern_europe/nl012/nl012.pdf	745	104	89	70
/en/publications/magazines/newsletter_eastern_europe/nl013/index.htm	158	49	43	22
/en/publications/magazines/newsletter_eastern_europe/nl013/nl013.pdf	206	42	77	63
/en/publications/magazines/newsletter_eastern_europe/nl014/index.htm	93		55	38
/en/publications/magazines/newsletter_eastern_europe/nl014/nl014.pdf	62		42	20
/en/publications/magazines/newsletter_eastern_europe/nl01s/index.htm	444	109	99	83
/en/publications/magazines/newsletter_eastern_europe/nl01s/nl_sh_2001.pdf	1202	494	246	166
/en/staff/as/index.htm	273	25	30	16
/gesis_aussenstelle/index.htm	14162	889	818	517
/gesis_aussenstelle/uebersicht.htm	1037	78	55	40
/index.htm	42793	4852	4874	3293
/information/foris/erhebung/index.htm	3537	676	1040	494
/information/foris/erhebung/preview/index.htm	766	148	158	96
/information/foris/index.htm	5139	631	806	456
/information/foris/recherche/index.htm	8081	1189	1229	642
/information/index.htm	16877	1723	1879	1139
/information/osteuropa/ineaste/index.htm	735	227	155	113
/information/osteuropa/proeaste/erhebung/index.shtm	160	26	16	5
/information/osteuropa/proeaste/index.htm	495	73	58	46
/information/osteuropa/proeaste/recherche/index.htm	207	31	30	22
/information/rechercheunterst/klassifikation/index.htm	2257	275	296	145
/information/rechercheunterst/klassifikation/klass.htm	1004	175	169	100
/information/rechercheunterst/klassifikation/klass.pdf	2247	474	327	226
/information/sofo/index.htm	5562	518	510	278
/information/sofo/recherche/index.htm	2168	146	102	41
/information/solis/index.htm	7228	771	857	551
/information/themen/europa.htm	730	121	104	65
/information/themen/fokus/balkan/balkan.pdf	1861	201	209	153
/information/themen/fokus/balkan/index.htm	653	34	78	42
/information/themen/fokus/bse/bse.pdf	1215	114	82	76
/information/themen/fokus/bse/index.htm	635	35	45	20
/information/themen/fokus/eu/eu.pdf	4493	614	531	408
/information/themen/fokus/eu/index.htm	964	50	168	117
/information/themen/fokus/euro/euro.pdf	168			168
/information/themen/fokus/euro/index.htm	152			152
/information/themen/fokus/gesundheit/gesundheit.pdf	4446	587	316	310
/information/themen/fokus/gesundheit/index.htm	573	37	38	32
/information/themen/fokus/h_gew/h_gew.pdf	25			25
/information/themen/fokus/h_gew/index.htm	41			41
/information/themen/fokus/islam/index.htm	2833	931	598	349
/information/themen/fokus/islam/islam.pdf	5307	1339	927	588
/information/themen/fokus/klima/index.htm	425	28	33	9
/information/themen/fokus/klima/klima.pdf	1550	150	199	172
/information/themen/fokus/pol_skan/index.htm	662	49	44	39
/information/themen/fokus/pol_skan/pol_skan.pdf	2344	296	250	206
/information/themen/fokus/rechtsradikalismus/index.htm	3419	276	233	200
/information/themen/fokus/rechtsradikalismus/rechtsradikalismus.pdf	5130	451	432	381
/information/themen/fokus/renten/index.htm	889	46	55	29
/information/themen/fokus/renten/renten.pdf	2703	266	137	129
/information/themen/fokus/school/index.htm	1590	146	317	191
/information/themen/fokus/school/school.pdf	5506	525	1009	726
/information/themen/fokus/solda/index.htm	593	34	44	31
/information/themen/fokus/solda/solda.pdf	1614	226	140	139
/information/themen/fokusplus/amerika/index.htm	7389	2294	2287	1165
/information/themen/fokusplus/index.htm	4164	1424	791	315

/information/zeitschriften/deutschspr/index.htm	2918	322	315	231
/information/zeitschriften/deutschspr/zeitschriften.doc	8	3	2	
/information/zeitschriften/deutschspr/zeitschriften.pdf	1176	261	145	132
/information/zeitschriften/deutschspr/zsliste_gesamt.htm	495	116	82	55
/information/zeitschriften/index.htm	2220	230	239	160
/information/zeitschriften/osteuropa/index.shtm	707	62	66	72
/iz/index.htm	35148	2676	2915	1883
/kooperation/information/osteuropa/associ.htm	339	38	35	27
/kooperation/information/osteuropa/index.htm	1157	59	53	46
/kooperation/information/osteuropa/presse.htm	433	38	87	41
/mitarbeiter/as/index.htm	1518	251	193	72
/veranstaltungen/transformationsforschung/index.htm	1216	75	70	47

5.3.4 Suchmaschinen-PageViews von Schlüssel-URLs

Die folgende Tabelle zeigt die Ergebnisse der Suchmaschinen-PageView-Auszählungen für die ausgewählten URLs als Ausschnitt aus der Zeitreihen-Analyse der Datei fils2001.log.

Die Tabelle zeigt folgende Spalten:

- URL,
- Summe der PageViews im Untersuchungszeitraum,
- PageViews pro Monat für die Monate Oktober, November und Dezember.

URL	Summe	2001-10	2001-11	2001-12
/bestellen/iz/index.htm	365	56	25	74
/en/cooperation/information/eastern_europe/associ.shtm	195	53	36	41
/en/cooperation/information/eastern_europe/index.htm	210	44	34	60
/en/cooperation/information/eastern_europe/network_ee.htm	88	29	16	21
/en/events/transformation/index.htm	193	57	39	45
/en/gesis_branch_office/index.htm	411	80	71	70
/en/gesis_branch_office/overview.htm	79	23	10	25
/en/information/eastern_europe/ineaste/index.htm	258	71	52	63
/en/information/eastern_europe/proeaste/index.htm	280	69	79	65
/en/information/eastern_europe/proeaste/search/index.htm	78	11	13	23
/en/information/eastern_europe/proeaste/survey/index.htm	78	13	10	23
/en/information/journals/eastern_europe/index.htm	373	73	82	73
/en/information/theme/europa.htm	96	18	11	27
/en/publications/magazines/newsletter_eastern_europe/index.htm	615	90	65	91
/en/publications/magazines/newsletter_eastern_europe/nl001/index.htm	326	68	74	48
/en/publications/magazines/newsletter_eastern_europe/nl001/nl001.pdf	99	7	29	32
/en/publications/magazines/newsletter_eastern_europe/nl002/index.htm	300	64	43	51
/en/publications/magazines/newsletter_eastern_europe/nl002/nl002.pdf	112	7	31	34
/en/publications/magazines/newsletter_eastern_europe/nl003/index.htm	346	66	52	79
/en/publications/magazines/newsletter_eastern_europe/nl003/nl003.pdf	115	8	37	32
/en/publications/magazines/newsletter_eastern_europe/nl004/index.htm	311	70	33	63
/en/publications/magazines/newsletter_eastern_europe/nl004/nl004.pdf	136	8	59	36
/en/publications/magazines/newsletter_eastern_europe/nl011/index.htm	295	66	45	48

/en/publications/magazines/newsletter_eastern_europe/nl011/nl011.pdf	113	7	36	36
/en/publications/magazines/newsletter_eastern_europe/nl012/index.htm	263	68	42	44
/en/publications/magazines/newsletter_eastern_europe/nl012/nl012.pdf	103	7	35	30
/en/publications/magazines/newsletter_eastern_europe/nl013/index.htm	202	51	46	73
/en/publications/magazines/newsletter_eastern_europe/nl013/nl013.pdf	70	6	31	32
/en/publications/magazines/newsletter_eastern_europe/nl014/index.htm	84		28	56
/en/publications/magazines/newsletter_eastern_europe/nl014/nl014.pdf	10		2	8
/en/publications/magazines/newsletter_eastern_europe/nl01s/index.htm	232	66	52	53
/en/publications/magazines/newsletter_eastern_europe/nl01s/nl_sh_2001.pdf	96	8	36	36
/en/staff/as/index.htm	180	40	24	23
/gesis_aussenstelle/index.htm	790	105	108	87
/gesis_aussenstelle/uebersicht.htm	301	47	41	49
/index.htm	3292	157	155	165
/information/foris/erhebung/index.htm	694	104	62	118
/information/foris/erhebung/preview/index.htm	436	46	30	86
/information/foris/index.htm	640	125	71	119
/information/foris/recherche/index.htm	459	74	33	72
/information/index.htm	776	105	79	108
/information/osteuropa/ineaste/index.htm	559	97	67	73
/information/osteuropa/proeaste/erhebung/index.shtm	167	23	14	33
/information/osteuropa/proeaste/index.htm	316	55	43	60
/information/osteuropa/proeaste/recherche/index.htm	159	21	17	41
/information/rechercheunterst/klassifikation/index.htm	272	45	22	65
/information/rechercheunterst/klassifikation/klass.htm	165	14	13	34
/information/rechercheunterst/klassifikation/klass.pdf	29	3	2	4
/information/sofo/index.htm	736	86	81	97
/information/sofo/recherche/index.htm	289	41	19	51
/information/solis/index.htm	556	83	57	179
/information/themen/europa.htm	332	68	25	71
/information/themen/fokus/balkan/balkan.pdf	80	9	5	35
/information/themen/fokus/balkan/index.htm	290	52	20	69
/information/themen/fokus/bse/bse.pdf	75	9	4	36
/information/themen/fokus/bse/index.htm	306	61	23	78
/information/themen/fokus/eu/eu.pdf	75	9	4	33
/information/themen/fokus/eu/index.htm	282	48	21	64
/information/themen/fokus/euro/euro.pdf	8			8
/information/themen/fokus/euro/index.htm	44			44
/information/themen/fokus/gesundheit/gesundheit.pdf	80	9	7	37
/information/themen/fokus/gesundheit/index.htm	310	66	25	74
/information/themen/fokus/h_gew/h_gew.pdf	2			2
/information/themen/fokus/h_gew/index.htm	22			22
/information/themen/fokus/islam/index.htm	367	71	46	98
/information/themen/fokus/islam/islam.pdf	79	9	5	36
/information/themen/fokus/klima/index.htm	304	52	28	78
/information/themen/fokus/klima/klima.pdf	74	8	4	35
/information/themen/fokus/pol_skan/index.htm	324	63	23	79
/information/themen/fokus/pol_skan/pol_skan.pdf	80	9	4	35
/information/themen/fokus/rechtsradikalismus/index.htm	313	58	24	81
/information/themen/fokus/rechtsradikalismus/rechtsradikalismus.pdf	79	9	4	32
/information/themen/fokus/renten/index.htm	315	55	27	74
/information/themen/fokus/renten/renten.pdf	81	9	4	38
/information/themen/fokus/school/index.htm	298	52	24	71
/information/themen/fokus/school/school.pdf	85	10	5	36
/information/themen/fokus/solda/index.htm	319	57	28	79
/information/themen/fokus/solda/solda.pdf	84	9	5	36
/information/themen/fokusplus/amerika/index.htm	177	49	26	57
/information/themen/fokusplus/index.htm	328	87	78	131
/information/zeitschriften/deutschspr/index.htm	397	52	23	134

/information/zeitschriften/deutschspr/zeitschriften.pdf	59	6	4	11
/information/zeitschriften/deutschspr/zsliste_gesamt.htm	176	20	16	32
/information/zeitschriften/index.htm	471	78	79	87
/information/zeitschriften/osteuropa/index.shtm	313	52	24	74
/iz/index.htm	1037	146	107	129
/kooperation/information/osteuropa/associ.htm	153	30	13	26
/kooperation/information/osteuropa/index.htm	299	58	69	40
/kooperation/information/osteuropa/presse.htm	237	42	66	26
/mitarbeiter/as/index.htm	358	62	77	53
/veranstaltungen/transformationsforschung/index.htm	369	65	76	62

5.4 PageViews: Top 20 URLs

Für einen Überblick über das Hauptinteresse der Nutzer und der Suchmaschinen-Schwerpunkte werden die URLs mit den ca. 20 meisten PageViews gelistet.

5.4.1 Besucher-Interesse gesamt (filb2001all.log)

Seite/Objekt	PageViews (PV)
/za/index.htm	52637
/zuma/index.htm	43769
/index.htm	42793
/iz/index.htm	35148
/information/index.htm	16877
/datenservice/index.htm	16299
/bestellen/index.htm	15733
/publikationen/index.htm	15338
/en/data_service/eurobarometer/animate.js	14625
/gesis_aussenstelle/index.htm	14162
/en/data_service/eurobarometer/index.htm	13098
/methodenberatung/index.htm	11383
/dauerbeobachtung/sozialindikatoren/index.htm	10635
/dauerbeobachtung/mikrodaten/index.htm	10620
/socioguide/index.htm	10266
/mitarbeiter/index.htm	9998
/software/index.htm	9977
/forschung/index.htm	9308

5.4.2 Besucher-Interesse Schlüssel-URLs (filb2001.log)

Seite/Objekt	PageViews (PV)
/index.htm	42793
/iz/index.htm	35148
/bestellen/iz/index.htm	22678
/information/index.htm	16877
/gesis_aussenstelle/index.htm	14162
/information/foris/recherche/index.htm	8081
/information/themen/fokusplus/amerika/index.htm	7389
/information/solis/index.htm	7228
/information/sofo/index.htm	5562
/information/themen/fokus/school/school.pdf	5506
/information/themen/fokus/islam/islam.pdf	5307
/information/foris/index.htm	5139
/information/themen/fokus/rechtsradikalismus/rechtsradikalismus.pdf	5130
/information/themen/fokus/eu/eu.pdf	4493
/information/themen/fokus/gesundheit/gesundheit.pdf	4446
/information/themen/fokusplus/index.htm	4164
/information/foris/erhebung/index.htm	3537
/information/themen/fokus/rechtsradikalismus/index.htm	3419
/information/zeitschriften/deutschspr/index.htm	2918

5.4.3 Suchmaschinen-Schwerpunkte gesamt (fils2001all.log)

Seite/Objekt	PageViews (PV)
/dauerbeobachtung/sozialindikatoren/index.htm	104636
/dauerbeobachtung/mikrodaten/index.htm	10633
/dauerbeobachtung/einkommen/index.htm	9595
/methodenberatung/zis/index.htm	5174
/dauerbeobachtung/allbus/index.htm	3573
/index.htm	3292
/software/index.htm	2735
/zuma/index.htm	2252
/methodenberatung/index.htm	2008
/methodenberatung/textanalyse/index.htm	1153
/publikationen/berichte/zuma_arbeitsberichte/index.htm	1141
/mitarbeiter/zuma/index.htm	1129
/methodenberatung/vercodung/index.htm	1127
/iz/index.htm	1037
/publikationen/zeitschriften/zuma_nachrichten/index.htm	1021
/za/index.htm	973
/en/data_service/eurobarometer/index.htm	972
/veranstaltungen/zuma/index.htm	954
/datenservice/allbus/index.htm	937

5.4.4 Suchmaschinen-Schwerpunkte Schlüssel-URLs (fils2001.log)

Seite/Objekt	PageViews (PV)
/index.htm	3292
/iz/index.htm	1037
/gesis_aussenstelle/index.htm	790
/information/index.htm	776
/information/sofo/index.htm	736
/information/foris/erhebung/index.htm	694
/information/foris/index.htm	640
/en/publications/magazines/newsletter_eastern_europe/index.htm	615
/information/osteuropa/ineaste/index.htm	559
/information/solis/index.htm	556
/information/zeitschriften/index.htm	471
/information/foris/recherche/index.htm	459
/information/foris/erhebung/preview/index.htm	436
/en/gesis_branch_office/index.htm	411
/information/zeitschriften/deutschspr/index.htm	397
/en/information/journals/eastern_europe/index.htm	373
/veranstaltungen/transformationsforschung/index.htm	369
/information/themen/fokus/islam/index.htm	367
/bestellen/iz/index.htm	365

5.5 Suchmaschinen im Detail

Mit den folgenden Beispielen soll auf weitere Auswertungsmöglichkeiten von WebSuxess, u.a. die differenzierte Referrer-Analysen, hingewiesen werden.

5.5.1 Suchmaschinen-Domänen auf dem Gesamt-Angebot

Die Filter-Datei fils2001all.log enthält alle gefilterten Suchmaschinen-Zugriffe auf das Gesamtangebot.

Die folgende Tabelle zeigt die Suchmaschinen-Domänen (als Hinweis auf die Suchmaschinen-Firma) und die zugehörigen Zugriffs-Häufigkeiten. Unbekannte Domänen gehören zu Such-Robotern, die eine IP-Adresse aus einem "verdächtigen" Subnetz besitzen, denen aber kein DNS-Name zugewiesen wurde.

Second-Level Domain	Besucher (= Anzahl Roboter)	Visits	PageViews
googlebot.com	125	57922	314899
unbekannt	30	3122	213172
fireball.de	4	7846	38872
Openfind.com	26	11902	29057
fastsearch.net	14	455	28441
inktomi.com	85	25751	27962
inktomisearch.com	26	11960	20150
alexa.com	25	2437	9073
av.com	14	509	8802
ne.jp	3	349	7259
speedfind.de	1	1536	3876
Tivra.com	1	374	3357
co.jp	1	992	1716
lycos.com	24	1008	1148
Summe			**707784**

Von diesen 707.784 Suchmaschinen-PageViews wurden (nur) 347.544 durch die WebSuxess-eigenen Suchmaschinen-Dateien erkannt:

Suchmaschine	Roboter	Visits	PageViews
Google	125	57922	314899
Inktomi	27	12011	20292
Fireball	1	5063	9496
Infoseek	1	992	1716
Lycos	23	1001	1141
Summe			**347544**

5.5.2 Suchmaschinen-Domänen auf Schlüssel-URLs

Im Vergleich zu den Suchmaschinen-Zugriffen auf das Gesamtangebot zeigt die folgende Tabelle die Suchmaschinen-Zugriffe auf die ausgewählten Schlüssel-URLs (Filter-Datei: fils2001.log):

Second-Level Domain	Besucher (= Anzahl Roboter)	Visits	PageViews
googlebot.com	120	6107	11983
unbekannt	26	2732	6766
fastsearch.net	13	941	1804
fireball.de	4	910	1488
inktomi.com	74	931	1204

alexa.com	24	482	645
openfind.com	24	573	626
ne.jp	3	184	495
inktomisearch.com	22	464	471
Tivra.com	1	176	406
av.com	9	118	226
co.jp	1	140	146
speedfind.de	1	104	142
lycos.com	9	28	28
Summe			**26430**

5.5.3 Ansatz: AdClick-Analyse

WebSuxess bietet die Möglichkeit, mit der "Kampagnenanalyse" zu untersuchen, von welchen (externen) Servern die Besucher auf das eigene Angebot verwiesen wurden. AdClicks sind Hyperlinks, die sich auf anderen Web-Seiten befinden. Mit einer solchen Untersuchung kann u.a. überprüft werden, von welchen Hosts aus besonders wirkungsvoll auf das eigene Angebot verwiesen wird.

Beispiel: Ausschnitt aus der AdClick-Server-Liste (WebSuxess, filb2001all.log)

Server	**AdClicks**
http://www.google.de	51675
http://www.social-science-gesis.de	34944
http://www.zuma-mannheim.de	8187
http://google.yahoo.com	6982
http://de.google.yahoo.com	6708
http://www.bonn.iz-soz.de	5705
http://www.berlin.iz-soz.de	2215
[unknown+origin]	2079
http://de.dir.yahoo.com	2060
http://www.google.ch	1987
http://www.dimdi.de	1859
http://www.issp.org	1853
http://mserv.rrzn.uni-hannover.de	1755
http://www.za.uni-koeln.de	1711
http://suche.lycos.de	1225
http://www.emnid.tnsofres.com	1033
http://de.altavista.com	955
http://jserv.rrzn.uni-hannover.de	778
http://search.de.altavista.com	756
http://www.bibliothek.uni-regensburg.de	751

http://search.netscape.com	738
http://www.altavista.com	735
http://www.data-archive.ac.uk	677
http://search-intl.netscape.com	614
http://www.google.com	585
http://www.or.zuma-mannheim.de	549
http://userpage.fu-berlin.de	530
http://bserv.rrzn.uni-hannover.de	482
http://www.zpid.de	424
http://193.175.238.12	364
http://www.infoseek.de	361
http://translate.google.com	331
http://www.forschungsgruppewahlen.de	322
http://wwwpsy.uni-muenster.de	318
http://www.mzes.uni-mannheim.de	312
http://www.google.co.uk	296
http://uk.google.yahoo.com	292
http://www.uni-duesseldorf.de	288
http://www.ask.com	279
http://www.nsd.uib.no	271
http://www.ew2.uni-mannheim.de	266
http://www.gsu.edu	244
http://www.lycos.de	243
http://www.sozialforschung.de	243
http://idw-online.de	231
http://www.google.fr	217
http://wissen.fireball.de	215
http://www.dezzktop.de	208
http://www.ccsd.ca	202
bookmarks	200
http://193.175.239.23	195
http://www.uni-jena.de	194
http://www2.hu-berlin.de	188
http://www.soz.uni-heidelberg.de	177
http://search.altavista.com	177
http://www.die-frankfurt.de	171
http://www2.rz.hu-berlin.de	165
http://www.uni-koeln.de	165
http://www.statistik-bund.de	163
http://www.soziologie.de	162
http://www.fsd.uta.fi	161
http://suche.freenet.de	158
http://www.heise.de	156
http://www.uni-koblenz.de	154
http://www.alltheweb.com	149
http://212.227.33.241	143
http://www.infoseek.co.jp	139
http://www.sowi.uni-mannheim.de	138
http://www.uni-bamberg.de	134
http://www.spss-buch.de	133
http://dbserv.rrzn.uni-hannover.de	130

http://www.uni-tuebingen.de	130
http://brisbane.t-online.de	129
http://suchen.aol.de	127
http://idw.tu-clausthal.de	123
http://www.politikerscreen.de	123
http://www.intext.de	122
http://staff-www.uni-marburg.de	120
http://images.google.com	118
http://www.caloweb.de	117
http://www.uni-leipzig.de	115
http://www.uni-konstanz.de	115
http://www.google.it	114
http://www.uni-muenster.de	111
http://ww.google.de	109
http://www.wu-wien.ac.at	108
http://www.perlentaucher.de	106
http://www.essex.ac.uk	106
http://hotbot.lycos.com	105
http://www.bis.uni-oldenburg.de	103
http://www.uni-giessen.de	101
http://www.sil.org	101
http://www.netcraft.com	100

Hinweise:

1. Für eine "Kampagnenanalyse" müssen in WebSuxess über die Profiloptionen die eigene(n) Web-Adresse(n) ausgefiltert werden.

2. Es kann in WebSuxess separat untersucht werden, welche AdClicks speziell von Suchmaschinen kommen, zu diesem Zweck müssen allerdings die Suchmaschinen-Dateien von WebSuxess entsprechend gepflegt werden. Da dies den Aufwand für die vorliegende Untersuchung übersteigt, wird in den folgenden Beispielen mit den vorhandenen Suchmaschinen-Tabelle von WebSuxess (s. Kap. 2.2.3.1) gearbeitet.

Die **folgenden Tabellen sind also nur als methodische Muster** zu betrachten.

Beispiel: Ausschnitt aus der Zusammenfassung der Suchmaschinen-Verweise (WebSuxess, filb2001all.log):

Suchmaschine	**Verweise**
Google	24717
Fireball	1484

Yahoo! Deutschland	1244
MSN Deutschland Web Search	1221
Lycos Deutschland	974
MSN Web Search	618
Web.de	534
Altavista	333
Infoseek Deutschland	304
Northern Light	190
Fast Search	117
Excite - Deutsche Ausgabe	112
Dogpile	81
GoTo.com	73
AOL NetFind Deutschland	55
Excite	49
MSN UK Web Search	37
Lycos	22
AOL.com Search	14
AllesKlar	12
Speedfind	12
Apollo7	11
Evreka	10

Beispiel: Ausschnitt der AdClicks aus Verweisen der Google-Suchen-Funktion:

Internetadresse (, von denen Ihre Besucher weitergeleitet worden sind)	**AdClicks**
Google Suche nach	527
Google Suche nach eurobarometer	305
Google Suche nach textanalyse	170
Google Suche nach zuma	153
Google Suche nach schildkröte	123
Google Suche nach gesis	113
Google Suche nach terror in amerika	102
Google Suche nach terror gegen amerika	101
Google Suche nach umfragen	99
Google Suche nach gewalt in der schule	88
Google Suche nach rechtsradikalismus	88
Google Suche nach archiving	65
Google Suche nach osterweiterung	63
Google Suche nach allbus	59
Google Suche nach stadtplan mannheim	57
Google Suche nach qbase	55
Google Suche nach terror	54
Google Suche nach social indicators research	52
Google Suche nach neujahrsgedichte	51
Google Suche nach gentrification	49
Google Suche nach empirische sozialforschung	48

Google Suche nach politbarometer	42
Google Suche nach solis	41
Google Suche nach textpack	35
Google Suche nach korrespondenzanalyse	34
Google Suche nach social indicators	33
Google Suche nach sozialwissenschaften	33
Google Suche nach computer assisted telephone interviewing	32
Google Suche nach flash eurobarometer	32
Google Suche nach einkommen	30
Google Suche nach foris	29
Google Suche nach cati	28
Google Suche nach wertewandel	28
Google Suche nach spss download	27
Google Suche nach clusteranalyse	26
Google Suche nach mikrozensus	26
Google Suche nach text analysis	26
Google Suche nach teaching in germany	25
Google Suche nach mannheim map	25
Google Suche nach archivierung	25
Google Suche nach organigram	24
Google Suche nach map of mannheim	23
Google Suche nach sociological topics	23
Google Suche nach social inequality	22
Google Suche nach delphi methode	22
Google Suche nach eurobarameter	22
Google Suche nach umweltzeichen	21
Google Suche nach delphi-methode	20
Google Suche nach host	20
Google Suche nach inhaltsanalyse	20

5.6 Anmerkungen zu den Ergebnissen

- Von den rund 5 Mio. PageViews des Jahres 2001 stammen
  rund 3,1 Mio. (62%) von externen "Besuchern",
  rund 0,7 Mio. (14%) von Suchmaschinen und
  rund 1,2 Mio. (24%) von (pflegenden) Zugriffen aus der GESIS.

- Es wurden 379 Suchmaschinen-Roboter beobachtet, die pro Roboter
  durchschnittlich 333 Visits / 335 Tage (also rund 1 x täglich)
  durchschnittlich 5,6 PageViews pro Visit
  absolvierten.

- Von den rund 58.000 insgesamt abgerufenen unterschiedlichen Seiten wurden knapp 7.000 (12%) von Suchmaschinen angesprochen.

- Die ausgewählten Schlüssel-URLs repräsentieren bei den "Besuchern" mit
  92 von 26.000 abgerufenen Seiten = 0,4% des Angebotes.
  Sie erzielen
  262.000 von 3.097.000 = 8,5% PageViews,
  durch
  63.500 von 175.400 = 36% der Besucher.
  Dies lässt auf eine gute Auswahl schließen.

- Bei den Suchmaschinen-Zugriffen auf Schlüssel-URLs sind es
  26.400 von 707.800 PageViews = 3,7% und
  331 von 379 = 87% der Roboter.
  D.h.: Das selektive Suchmaschinen-Interesse an den Schlüssel-URLs ist im Vergleich zum Besucher-Interesse deutlich geringer.

- Zugriffe der "Besucher" auf URLs, die ein gezieltes fachliches Interesse vermuten lassen (unter Ausschluss der typischen Einstiegsseiten), liegen in der Spitzengruppe der Zugriffe bei 3-8.000 PageViews:

Seite/Objekt	PageViews (PV)
/information/foris/recherche/index.htm	8081
/information/themen/fokusplus/amerika/index.htm	7389
/information/solis/index.htm	7228
/information/sofo/index.htm	5562
/information/themen/fokus/school/school.pdf	5506
/information/themen/fokus/islam/islam.pdf	5307
/information/foris/index.htm	5139
/information/themen/fokus/rechtsradikalismus/rechtsradikalismus.pdf	5130
/information/themen/fokus/eu/eu.pdf	4493
/information/themen/fokus/gesundheit/gesundheit.pdf	4446
/information/themen/fokusplus/index.htm	4164
/information/foris/erhebung/index.htm	3537
/information/themen/fokus/rechtsradikalismus/index.htm	3419
/information/zeitschriften/deutschspr/index.htm	2918

- Bei der Bewertung der PageViews durch "Besucher" ist zu berücksichtigen, dass über 80% der Client-IP-Adressen nur je einmal im Auswertungszeitraum registriert wurden. Dies gilt sowohl für das Gesamtangebot als auch die Schlüssel-URLs.

- Das Interesse der verschiedenen Suchmaschinen-Anbieter am GESIS-Angebot (gemessen an den PageViews) ist sehr unterschiedlich

300.000 PV bei Google
1.000 PV bei Lycos
und kann ggf. durch explizite Registrierung gefördert werden.

- Eine Arabeske ist die Beobachtung, dass nach der AdClick-Analyse die Suchmaschine Google rund 52.000 PageViews auf unser Angebot erzeugte, dem ein ein Analyseaufwand von rund 315.000 PageViews der Suchroboter dieser Firma gegenüber steht.

6 Produktionsverfahren 2002

6.1 Implementiertes Verfahren

Auf der Grundlage der dargestellten Verfahrens-Empfehlungen und unter Berücksichtigung ergänzender Auswertungsbedürfnisse wurde folgendes Produktionsverfahren implementiert:

1. Die Log-Daten des Servers werden täglich kurz nach Mitternacht mit Hilfe einer Perl-Prozedur (LogBearbeitung.pl) gefiltert, dabei entstehen folgende Dateien, die auf einem zentralen Server in Berlin abgelegt werden:
   - Gesamt-WWW-Log: original.log
   - gefilterte Navigationselemente: sonder.log
   - Suchmaschinen (ohne Navigationselemente): suchm.log
   - GESIS-Zugriffe (ohne Navigationselemente): gesis.log
   - Besucher-Zugriffe (ohne Navigation, Suchmaschinen, GESIS): extend1.log

2. Die Datei extend1.log wird täglich, die übrigen Dateien werden zu Beginn eines jeden Monats auf einen zentralen Server nach Bonn kopiert.

3. Täglich werden in Berlin über die Dateien extend1.log und gesis.log Scheduler-gesteuert WebSuxess-Auswertungen für ca. 45 Profile durchgeführt und zentral unter http://193.175.239.64/statistik als tagesaktuelle HTML-Berichte GESIS-öffentlich bereitgestellt.

4. Zu Beginn eines jeden Monats werden aus den Dateien extend1.log und suchm.log die von den Fachabteilungen vorgegebenen Schlüssel-URLs gefiltert und in die Dateien filb2002.log bzw. fils2002.log ge-

speichert. Diese werden als monatliche Zeitreihe aufbereitet und als Berichts-Dateien zeitreihe2002-b.csv bzw. zeitreihe2002-s.csv IZ-öffentlich bereitgestellt.

5. Parallel hierzu wird in Bonn per WinInstall eine vorkonfigurierte Version von WebSuxess4 zur lokalen Installation auf den Arbeitsplätzen der Mitarbeiter bereitgestellt, mit Auswertungsprofilen für die in Bonn gespeicherten Dateien original.log, sonder.log, suchm.log, gesis.log, extend1.log.
   Hierdurch werden individuelle tagesaktuelle Untersuchungen der Benutzerzugriffe auf die Datei extend1.log sowie monatliche Analysen zu Sonderfragen sowohl über die Gesamtdatei als auch über die Selektionen ermöglicht.

6.2 Filterregeln

Die tägliche Filterung der Log-Datei erfogt nach fogenden Regeln:

1. Es werden alle Datensätze mit Zugriffen auf URLs mit folgenden Namensteilen entfernt und in der Datei sonder.log gespeichert:
   .gif
   .css
   .js
   .dll
   .jpg
   .ico
   robots.txt

2. Es werden alle Datensätze mit Zugriffen von einer GESIS-IP-Adresse entfernt und in der Datei gesis.log gespeichert.

193.175.238.*	IZ Bonn
193.175.239.*	IZ Berlin
193.196.10.*	ZUMA Mannheim
134.155.33.*	Proxy-Server RZ Uni Mannheim
134.95.45.*	ZA Köln

3. Es werden alle Datensätze mit Zugriffen von einer Suchmaschinen-IP-Adresse entfernt und in der Datei suchm.log gespeichert.

6.2.1 Dokumentation des Skript LogBearbeitung.pl

```
$Eingabeverzeichnis="c:/logfiles/gesis/OriginalLog/W3SVC1";
$SteuerEingabeverzeichnis="c:/logfiles/gesis/Parameter";
$Ausgabeverzeichnis="c:/logfiles/gesis/w3svc1";
$Protokollverzeichnis="c:/logfiles/gesis/Protokoll";
$fileOriginal=">> $Ausgabeverzeichnis/Original.log";
$fileSonder=">> $Ausgabeverzeichnis/Sonder.log";
$fileGesis=">> $Ausgabeverzeichnis/Gesis.log";
$fileSuchm=">> $Ausgabeverzeichnis/Suchm.log";
$fileRest=">> $Ausgabeverzeichnis/extend1.log";
$fileProtokoll=">> $Protokollverzeichnis/Protokoll.log";
$fileGesisURL="$SteuerEingabeverzeichnis/gesisurl.txt";
$fileSuchmaURL="$SteuerEingabeverzeichnis/suchmaurl.txt";

($sec,$min,$hour,$mday,$mon,$year,$wday,$yday,$isdst) = localtime(time);
$year += 1900;
$year = sprintf("%02d", $year % 100);
$mon += 1;
$mon = sprintf("%02d", $mon % 100);
$mday = sprintf("%02d", $mday % 100);
$fileakt="ex$year$mon$mday.log";

open(OUTProtokoll, $fileProtokoll) || die "Can't open Protokoll.log: $!\n";
print  OUTProtokoll "*-------------------------------------------------------*\n";
print  OUTProtokoll "Protokoll vom: $mday.$mon.$year um $hour:$min\n";
open(OUTOriginal, $fileOriginal) || (die "!!!!!!!!!!!!!!!!Can't open Original.log: $!\n" &&
print OUTProtokoll "!!!!!!!!!!!!!!!!Can't open Original.log: $!\n");
open(OUTSonder, $fileSonder) || (die "!!!!!!!!!!!!!!!!Can't open Sonder.log: $!\n" && print
OUTProtokoll "!!!!!!!!!!!!!!!!Can't open Sonder.log: $!\n");
open(OUTGesis, $fileGesis) || (die "!!!!!!!!!!!!!!!!Can't open Gesis.log: $!\n" && print
OUTProtokoll "!!!!!!!!!!!!!!!!Can't open Gesis.log: $!\n");
open(OUTSuchm, $fileSuchm) || (die "!!!!!!!!!!!!!!!!Can't open Suchm.log: $!\n" && print
OUTProtokoll "!!!!!!!!!!!!!!!!Can't open Suchm.log: $!\n");
open(OUTRest, $fileRest) || (die "!!!!!!!!!!!!!!!!Can't open Rest.log: $!\n" && print OUT-
Protokoll "!!!!!!!!!!!!!!!!Can't open Rest.log: $!\n");

open(INGesisURL, $fileGesisURL) || (die "!!!!!!!!!!!!!!!!Can't open gesisurl.txt: $!\n" &&
print OUTProtokoll "!!!!!!!!!!!!!!!!Can't open gesisurl.txt $!\n");
  while (<INGesisURL>) {
         if(/(([0-9]*\.)*[0-9]*)/){
               $gesis{$1}=1;
         }
  }

close INGesisURL;
open(INSuchmaURL, $fileSuchmaURL) || (die "!!!!!!!!!!!!!!!!Can't open gesisurl.txt: $!\n"
&& print OUTProtokoll "!!!!!!!!!!!!!!!!Can't open gesisurl.txt $!\n");
  while (<INSuchmaURL>) {
         if(/(([0-9]*\.)*[0-9]*)/){
               $suchma{$1}=1;
         }
  }
close INSuchmaURL;

chdir ($Eingabeverzeichnis);
@files = <*.log>;
$i = 0;
foreach $name (@files){
    next if (-d $name);
        if ($name !~ (/$fileakt/)){
               open(IN, "$Eingabeverzeichnis/$name") || (die "!!!!!!!!!!!!!!!!Can't open
$name: $!\n" && print OUTProtokoll "!!!!!!!!!!!!!!!!Can't open $name $!\n");
               while (<IN>) {
                      print OUTOriginal lc $_;
                      if(/\.gif|\.css|\.js|\.dll|\/robots.txt|\.jpg|\.ico/oi){
                             print OUTSonder lc $_;
                             next;
                      }
                      if (/(([0-9]*\.[0-9]*\.[0-9]*)\.[0-9]*)/){
                             if ($gesis{$1} || $gesis{$2}){
                                    print OUTGesis lc $_;
                                    next;
                             }
```

```
                                if ($suchma{$1} || $suchma{$2}){
                                        print OUTSuchm lc $_;
                                        next;
                                }
                        }
                if (/^#/){
                        print OUTRest $_;
                }else{
                        print OUTRest lc $_;
                        }
                }
                close IN;
                $i++;
                rename "$Eingabeverzeichnis/$name","$Eingabeverzeichnis/$name\_$i";
                print  OUTProtokoll "Datei: $name bearbeitet\n";
        }
}


close OUTOriginal;
close OUTSonder;
close OUTGesis;
close OUTSuchm;
close OUTRest;
close OUTProtokoll;
```

6.2.2 Dokumentation der Suchmaschinenliste

Die folgende Liste enthält den Erkenntnisstand Juni 2002 zu Suchmaschinen und Suchmaschinen-Gruppen, die das GESIS-Angebot besuchen. Die IP-Adressen werden entweder als 4-Byte-Adresse oder als 3-Byte-Adresse angegeben, mit folgender Bedeutung:

- 4-Byte-Adresse: exakt diese Adresse
- 3-Byte-Adresse: alle 255 Adressen in diesem Subnetz

```
129.250.233.33
130.75.2.24
142.75.65.141
144.140.254.227
149.174.106.123
149.174.34.13
151.189.12.147
166.90.205.55
171.64.75.83
192.215.220.3
192.41.33.220
192.49.214.115
193.102.192.185
193.14.34.204
193.189.238.2
193.189.238.7
193.252.117.42
193.252.117.43
193.7.255
193.7.255.189
193.7.255.22
194.126.132
194.168.54.50
```

```
194.168.55.12
194.221.102.128
194.221.102.137
194.224.53.173
194.231.30
194.232.15.65
194.239.250.89
194.251.241.160
194.45.170.244
194.64.51.100
194.64.51.231
194.65.155.68
194.65.79.190
194.7.1.40
194.97.50.196
194.97.50.203
194.97.8.163
195.101.94
195.101.94.208
195.110.96.106
195.121.7.215
195.145.119.23
195.170.70.244
```

195.186.3.242
195.210.91.189
195.235.51.227
195.248.63.82
195.3.97.3
195.58.160.40
195.58.177.120
195.63.60.109
195.63.81.16
195.92.249.46
198.161.157.212
198.3.103.205
199.172.146.105
199.172.146.99
199.172.148.11
200.231.206.38
200.231.206.53
202.1.232.102
202.1.238.114
202.186.13.86
202.214.130.2
202.3.13.16
202.3.13.17
202.3.14.151
202.3.14.157
203.108.8.12
203.109.252.72
203.164.2.108
203.89.227.140
204.187.152.26
204.202.132.19
204.202.140.215
205.180.85.19
205.188.133.5
205.188.180.249
205.188.180.25
206.129.0.232
206.253.217.38
206.3.4.200
206.46.189.11
206.79.171
206.79.171.196
206.79.171.51
206.79.171.54
207.138.42.25
207.138.42.32
207.140.168
207.156.250.35
207.200.81.135
207.200.83.135
207.46.238.120
207.68.176.250
207.68.176.254
207.68.185.57
208.185.214.132
208.185.214.133
208.185.214.134
208.233.51.3
208.237.254.40
208.254.3.130
208.45.133.15
209.10.180.30
209.123.16.9
209.132.193.8
209.185.143
209.202.148
209.202.192.35
209.202.193
209.202.195.206
209.247.208.41
209.247.208.44
209.247.40
209.67.227.90
209.67.229
209.73.162
209.73.164
209.73.180
210.150.10
210.150.160.234
210.150.25.37
211.169.241.210
211.18.214
211.32.119.136
211.72.252.184
212.123.93.125
212.166.67.6
212.172.247.162
212.20.160.80
212.227.116.121
212.227.63.58
212.48.4
212.72.39.219
212.74.101.10
213.168.94.36
213.198.31.161
213.208.133.110
213.36.100.185
213.69.45.45
216.115.102.243
216.115.106.163
216.115.109.6
216.115.109.7
216.136.130.252
216.136.227.113
216.167.36.217
216.218.223.17
216.239.33.101
216.239.39.101
216.239.46
216.248.193.241
216.248.199.88
216.27.131.50
216.32.86.230
216.34.102.218
216.34.102.230
216.34.197.132
216.35.103
216.35.116
216.35.194
216.35.70
216.40.246.10
216.86.229.28
217.115.138.235
217.12.3
217.72.195
62.104.23.50

62.119.21
62.12.134.34
62.144.160.2
62.144.211.70
62.144.211.71
62.144.98.240
62.154.244.102
62.172.197.150
62.41.154
62.52.160
62.67.200.53
63.251.4
64.12.149.245
64.12.153.184
64.124.237.128
64.124.237.146
64.14.53.148
64.14.53.254
64.14.85.172
64.15.129
64.15.202.140
64.15.202.151
64.15.227.200
64.15.227.203
64.152.75
64.158.138.26
64.241.242.219
64.244.109.25
64.58.76.178
64.58.77
64.68.72.17
64.68.82
64.68.86
64.7.208.39
64.70.201.41
65.214.36
65.214.36.242
65.214.39.11
65.214.39.7
65.214.39.8
66.196.73
66.51.203.10
66.7.131
66.77.73
66.77.74
66.77.74.20
66.77.74.21
80.69.224.33
80.78.233.188

6.3 Liste der täglichen WebSuxess-Auwertung

Die folgende Tabelle zeigt das Inhaltsverzeichnis der täglich auf der Grundlage der Datei extend1.log als html-Bericht (http://193.175.239.64/statistik) bereitgestellten Auswertung
(Stand: Juni 2002):

Auswertungszeitraum: 2002, 01.01.-31.12.	**Archiv 2001**
GESIS Angebot deutsch alles; alles **nur GESIS-Mitarbeiter** (Grundlage: gesis.log)	**GESIS Angebot english alles**
Literatur- & Forschungsinformation Datenbank SOLIS; Datenbank FORIS; Host-Zugänge; CD-ROM WISO III; Datenbank PROEastE; Datenbank SOFO; Datenbank INEastE; Zeitschriften; Rechercheunterstützung; Themenorientierte Angebote (wie FOKUS); Auftragsrecherchen; Beratung	**Literature & Research Information** Database SOLIS; Database FORIS; Access via Hosts; CD-ROM WISO III; Database PROEastE; Database SOFO; Database INEastE; Journals; Support for Searches; Thematically prepared Offers; Search Service; Consultation
Datenservice & Archivierung Suche: Daten & Dokumentation; ALLBUS; DDR - Neue Bundesländer; Politbarometer; Wahlstudien; Eurobarometer; ISSP; Studien aus Osteuropa; Historische Sozialforschung; Thematische Datenpools; Nutzerberatung; Bestellen & Downloads	**Data Service & Archiving** Search Data & Documentation; ALLBUS; GDR & new Federal States; Eurobarometer; ISSP; Eastern Europe; Consultation; Order & Downloads
Dauerbeobachtung ALLBUS; ISSP; Microdaten; Einkommen & Verbrauch; Soziale Indikatoren	**Social Monitoring** ALLBUS; ISSP; Microdata; Income & Expenditure; Social Indicators
Methodenberatung Datenerhebung; Stichprobe & Datenauswertung; Vercodung; Textanalyse; Telephonumfragen; Elektronisches Handbuch; Demographische Standards; Antragsformular; Mitarbeiter & Adressen	**Methods Consultation** Fielding; CATI; Sampling & Data-Analysis; Encoding; Text-Analysis; ZUMA Information System ZIS; Staff & Addresses

Forschung & Entwicklung Informationstechnologie; EUROLAB; Historische Sozialforschung; Dauerbeobachtung; Methodenforschung; Soziologische Themen	**Research & Development** Information Technology; EUROLAB; Methods Research; Historical Social Research; Sociological Topics
Software Clustan / ClustanGraphics; NSDstat Pro; Schildkröte	**Software** Badason; TEXTPACK
Publikationen Zeitschriften & Newsletter; Arbeits- & Forschungsberichte; Monographien & Sammelwerke; Zeitschriftenaufsätze	**Publications** ZA Information; IZ Telegram; Newsletter Eastern Europe; IZ Working Papers; Monographs & Compilations (IZ, ZA)
Bestellen & Downloads Literatur- & Forschungsinformation; Datenservice & Archivierung; Dauerbeobachtung; Methodenberatung; Forschung & Entwicklung; Software; Publikationen	**Order & Downloads** Data Service
Veranstaltungen National & International; Transformationsforschung; IZ Bonn; ZA Köln; ZUMA Mannheim; GESIS Außenstelle Berlin	
GESIS-Bibliotheken Empirische Sozialforschung; Informationstransfer Osteuropa	**GESIS Libraries** Empirical Social Research
Linksammlung SocioGuide Westeuropa ; Osteuropa ; spezielle SocioGuides	Link Collection **SocioGuide**
Kooperationen Literatur- & Forschungsinformation; Datenservice & Archivierung; Dauerbeobachtung; Forschung & Entwicklung	**Cooperation** Data Service & Archiving; Research & Development
Beratung Literatur- & Forschungsinformation; Datenservice & Archivierung; Dauerbeobachtung; Methodenberatung	**Consultation** Data Service & Archiving
Mitarbeiter & Adressen IZ Bonn; ZA Köln; ZUMA Mannheim; GESIS Außenstelle Berlin	**Staff & Addresses** IZ Bonn; ZA Cologne
Organisation GESIS im Überblick; IZ Bonn; ZA Köln; ZUMA Mannheim; GESIS Außenstelle Berlin	**Organization** IZ Bonn; ZA Cologne
Homepages IZ, ZA, ZUMA, GESIS-Außenstelle	**Homepages**

historische Analysen

- Server IZ Bonn + Berlin
- externe Verlinkung der GESIS-Homepage deutsch und englisch

6.4 Filterung der Schlüssel-URLs

Die zu selektierenden Schlüssel-URLs werden in der Filter-Datei filter-url.txt wie folgt gekennzeichnet:

- Jeder Record in filter-url.txt bezeichnet - ohne die Steuerzeichen # am Beginn / Ende des Records - eine URL im Feld cs-uri-stem,
- beginnt der Record mit dem Steurzeichen #, wird auf Gleichheit eines Substrings des Feldes ab Feldbeginn getestet,

- endet der Record mit dem Steuerzeichen #, wird auf Gleichheit eines Substrings des Feldes bis Feldende getestet,
- beginnt der Record mit dem Steuerzeichen # und endet der Record nicht mit dem Steuerzeichen #, wird darauf getestet, ob der Record gleich einem Substring des Feldes ab Anfang des Feldes ist.

Mit Hilfe der Datei filter-out.txt können nach den gleichen Regeln Substrings definiert werden, die von der Selektion explizit ausgeschlossen werden.

Es werden nur Datensätze selektiert, die einen Status-Code 200, 206 oder 306 besitzen.

6.4.1 Dokumentation des Skript filter.awk

```
BEGIN      {while ((getline < "filter-url.txt") > 0) {filter[$0]=$0}
           while ((getline < "filter-out.txt") > 0) {filterout[$0]=$0}
           code1 = "200"
           code2 = "206"
           code3 = "304"
           }
           {inrec=inrec+1
            # print (inrec "   " outrec  "    " $8)
            if (substr($1,1,1)=="\#") {
              print >> "filb.log"
              }
            else
              {if (($8!~code1)&&($8!~code2)&&($8!~code3)) {
                 print >> "fil-code.log"
                 next
                 }
               $7 = "-"
               $11= "-"
               $0=tolower($0)
               yyy="\#" $6 "\#"
               for (i in filter) {
                 zzz=filter[i]
                 if (yyy ~ zzz) {
                    for (j in filterout) {
                       out=filterout[j]
                       if (yyy ~ out) {
                          print >> "fil-out.log"
                          next
                          }
                    }
                    print >> "filb.log"
                    outrec=outrec+1
                    next
                    }
                 else {
                    print >> "filr.log"
                    }
               }
           }
```

6.4.2 Filter-Liste der Schlüssel-URLs

Datei: filter-url.txt

Stand Juni 2002

```
#/bestellen/iz/index.htm#
#/en/cooperation/information/eastern_europe/index.htm#
#/en/cooperation/information/eastern_europe/associ.shtm#
#/en/cooperation/information/eastern_europe/network_ee.htm#
#/en/events/transformation/index.htm#
#/en/gesis_branch_office/index.htm#
#/en/gesis_branch_office/overview.htm#
#/en/information/eastern_europe/ineaste/index.htm#
#/en/information/eastern_europe/proeaste/index.htm#
#/en/information/eastern_europe/proeaste/search/index.htm#
#/en/information/eastern_europe/proeaste/survey/index.htm#
#/en/information/foris/index.htm#
#/en/information/foris/search/index.htm#
#/en/information/foris/survey/index.htm#
#/en/information/journals/index.htm#
#/en/information/journals/eastern_europe/index.htm#
#/en/information/sofo/index.htm#
#/en/information/solis/index.htm#
#/en/information/support/classification/index.htm#
#/en/information/support/methods/index.htm#
#/en/information/theme/europa.htm#
#/en/publications/magazines/newsletter_eastern_europe/index.htm#
#/en/publications/magazines/newsletter_eastern_europe/nl014/index.htm#
#/en/publications/magazines/newsletter_eastern_europe/nl014/nl014.pdf#
#/en/publications/magazines/newsletter_eastern_europe/nl013/index.htm#
#/en/publications/magazines/newsletter_eastern_europe/nl013/nl013.pdf#
#/en/publications/magazines/newsletter_eastern_europe/nl01s/index.htm#
#/en/publications/magazines/newsletter_eastern_europe/nl01s/nl_sh_2001.pdf#
#/en/publications/magazines/newsletter_eastern_europe/nl012/index.htm#
#/en/publications/magazines/newsletter_eastern_europe/nl012/nl012.pdf#
#/en/publications/magazines/newsletter_eastern_europe/nl011/index.htm#
#/en/publications/magazines/newsletter_eastern_europe/nl011/nl011.pdf#
#/en/publications/magazines/newsletter_eastern_europe/nl004/index.htm#
#/en/publications/magazines/newsletter_eastern_europe/nl004/nl004.pdf#
#/en/publications/magazines/newsletter_eastern_europe/nl003/index.htm#
#/en/publications/magazines/newsletter_eastern_europe/nl003/nl003.pdf#
#/en/publications/magazines/newsletter_eastern_europe/nl002/index.htm#
#/en/publications/magazines/newsletter_eastern_europe/nl002/nl002.pdf#
#/en/publications/magazines/newsletter_eastern_europe/nl001/index.htm#
#/en/publications/magazines/newsletter_eastern_europe/nl001/nl001.pdf#
#/en/publications/magazines/newsletter_eastern_europe/archiv/index.htm#
#/en/staff/as/index.htm#
#/gesis_aussenstelle/index.htm#
#/gesis_aussenstelle/uebersicht.htm#
#/index.htm#
#/iz/index.htm#
#/information/index.htm#
#/information/foris/erhebung/index.htm#
#/information/foris/erhebung/preview/index.htm#
#/information/foris/index.htm#
#/information/foris/recherche/index.htm#
#/forschung/informationstechnologie/index.htm#
#/forschung/informationstechnologie/informationssysteme.htm#
#/forschung/informationstechnologie/heterogenitaet.htm#
#/forschung/informationstechnologie/softwareergonomie.htm#
#/forschung/informationstechnologie/datenweiterverarbeitung.htm#
#/forschung/informationstechnologie/evaluation.htm#
#/information/osteuropa/ineaste/index.htm#
#/information/osteuropa/ineaste/index.htm#
#/information/osteuropa/proeaste/erhebung/index.shtm#
#/information/osteuropa/index.htm#
#/information/osteuropa/proeaste/recherche/index.htm#
#/information/recherche/host.htm#
#/information/recherche/kosten.htm#
#/information/recherche/cd.htm#
#/information/rechercheunterst/klassifikation/index.htm#
```

```
#/information/rechercheunterst/index.htm#
#/information/rechercheunterst/klassifikation/klass.htm#
#/information/rechercheunterst/klassifikation/klass.pdf
#/information/rechercheunterst/methodenliste/index.htm#
#/information/sofo/index.htm#
#/information/sofo/recherche/index.htm#
#/information/solis/index.htm#
#/information/solis/origb.htm#
#/information/themen/europa.htm#
#/information/themen/fokus/balkan/index.htm#
#/information/themen/fokus/balkan/balkan.pdf#
#/information/themen/fokus/bse/index.htm#
#/information/themen/fokus/bse/bse.pdf#
#/information/themen/fokus/eu/index.htm#
#/information/themen/fokus/eu/eu.pdf#
#/information/themen/fokus/euro/index.htm#
#/information/themen/fokus/euro/euro.pdf#
#/information/themen/fokus/gesundheit/index.htm#
#/information/themen/fokus/gesundheit/gesundheit.pdf#
#/information/themen/fokus/h_gew/index.htm#
#/information/themen/fokus/h_gew/h_gew.pdf#
#/information/themen/fokus/islam/index.htm#
#/information/themen/fokus/islam/islam.pdf#
#/information/themen/fokus/klima/index.htm#
#/information/themen/fokus/klima/klima.pdf#
#/information/themen/fokus/pol_skan/index.htm#
#/information/themen/fokus/pol_skan/pol_skan.pdf#
#/information/themen/fokus/rechtsradikalismus/index.htm#
#/information/themen/fokus/rechtsradikalismus/rechtsradikalismus.pdf#
#/information/themen/fokus/renten/index.htm#
#/information/themen/fokus/renten/renten.pdf#
#/information/themen/fokus/school/index.htm#
#/information/themen/fokus/school/school.pdf#
#/information/themen/fokus/solda/index.htm#
#/information/themen/fokus/solda/solda.pdf#
#/information/themen/fokusplus/index.htm#
#/information/themen/fokusplus/amerika/index.htm#
#/information/themen/fokusplus/israel/index.htm#
#/information/themen/fokusplus/schule&gewalt/index.htm#
#/information/themen/sofid/index.htm#
#/information/zeitschriften/deutschspr/index.htm#
#/information/zeitschriften/deutschspr/zeitschriften.doc#
#/information/zeitschriften/deutschspr/zeitschriften.pdf#
#/information/zeitschriften/deutschspr/zsliste_gesamt.htm#
#/information/zeitschriften/index.htm#
#/information/zeitschriften/osteuropa/index.shtm#
#/kooperation/information/osteuropa/associ.htm#
#/kooperation/information/osteuropa/index.htm#
#/kooperation/information/osteuropa/presse.htm#
#/mitarbeiter/as/index.htm#
#/publikationen/berichte/iz_arbeitsberichte/index.htm#
#/publikationen/berichte/iz_arbeitsberichte/pdf/
#/publicationen/zeitschriften/newsletter_osteuropa/index.htm#
#/socioguide/index.htm#
#/veranstaltungen/index.htm#
#/veranstaltungen/national_international/index.htm#
#/veranstaltungen/transformationsforschung/index.htm#
```

7 Zusammenfassung

Zweck der vorliegenden Studie ist die Evaluation aussagefähiger Nutzungs-Kennzahlen aus den Log-Dateien von WWW-Servern am Beispiel der Log-Ergebnisse des IIS-Servers www.gesis.org für die Jahre 2001 und 2002.

In Kap. 2 werden zunächst die prinzipiellen Eigenschaften der Log-Daten systematisch untersucht, u.a. die Problematik der Identifikation der Besucher anhand der IP-Adresse (DHCP, Proxy), die Behandlung von Suchmaschinen, die Unterscheidung von „Hits“, „PageViews“ und „Visits“, die Identifikation der angewählten URLs (Gross-/Kleinschreibung!), Befehlstypen und Statuscodes.

Aufgrund der in Kapitel 2.2 beschriebenen Randbedingungen wird empfohlen, die routinemäßige Ermittlung von Kennzahlen wie folgt einzugrenzen:

- Für die operative Nutzungs-Analyse der „Benutzer-Zugriffe“ werden aus den Log-Dateien alle Zugriffe
  a) auf Navigationselemente
     (URLs mit .gif, .css, .js, .dll, .jpg, .ico)
  b) von Suchmaschinen
     (anhand einer gepflegten Suchmaschinenliste),
  c) von IP-Adressen des Anbieters
     (hier: GESIS-Institute)

  herausgefiltert, in getrennten Dateien gespeichert und bei Bedarf separat ausgewertet.
- Es werden nur Datensätze mit den Statuscodes (sc-status) 200, 206 und 304 berücksichtigt.
- Zugriffe von Proxy-Servern werden als Indiz für triggernde Benutzeranfragen unverändert berücksichtigt.
- Als zentrale Kennzahl der Nutzung wird die Variable
  Anzahl PageViews pro Schlüssel-URL pro Zeiteinheit
  verwendet.
  Schlüssel-URL ist dabei ein einzelnes, vom Client anklickbares Element, welches im Feld cs-uri-stem des WWW-Logs aufgezeichnet wird. Schlüssel-URLs werden z.B. von den Fachabteilungen als repräsentative Seiten für die Beurteilung von Nutzungen benannt.

Als Zeiteinheit wird das Monatsraster der Teile "Jahr" und "Monat" des Datum-Feldes des WWW-Logs empfohlen.

- Mit Ausnahme einer PageView-Summe über alle Inhalts-Seiten des Angebotes wird auf die Ermittlung von Page-Views für Gruppen von Seiten verzichtet, da diese Zahlen nur eine geringe vergleichende Relevanz besitzen.
- Es wird wegen der in Kap. 2.2 geschilderten Probleme darauf verzichtet, Besucher oder Besuchergruppen zu individualisieren.

Es wird darauf hingewiesen, dass die Umsetzung dieser Empfehlungen den Einsatz von Filterprozeduren erfordert, da Tools wie WebSuxess diese Anforderungen nicht erfüllen. Ein Satz von Prototypen solcher Prozeduren wird in Kap. 3 beschrieben und untersucht. Die Produktionsversionen dieser Filterprozeduren (Stand: Juli 2002) sind in Kap. 6 dokumentiert und stehen Interessenten auf Anforderung zur Verfügung (s. Kap. 8 Anhang).

Kap. 4 beschreibt eine Verfahrensempfehlung für regelmäßige Routine-Auswertungen auf der Grundlage von Filterprozeduren einschließlich der Erstellung von Zeitreihen für Schlüssel-URLs.

In den Kapiteln 4.3 und 5 sind exemplarische Ergebnisse der Auswertung der Log-Dateien des Servers www.gesis.org zusammengestellt.

Es wird auf die Mengenverhältnisse der in den Log-Dateien dokumentierten unterschiedlichen Zugriffstypen hingewiesen.
Als Beispiel: Die Ergebnisdateien für den Zeitraum 1.1.-13.6.02 hatten folgende Größen (proportional zu der Anzahl der Records):

• Gesamtdatei:	1.586 MB,	rund 10 Mio. Hits rund 0,6 Mio. Visits
• Navigationselemente:	1.214 MB	(77 % von Gesamt)
• Rest:	372 MB	(23 % von Gesamt)
davon		
Suchmaschinen:	72 MB	(20 % von Rest)
GESIS:	46 MB	(12 % von Rest)
Besucher:	254 MB	(68 % von Rest, 16 % von Gesamt (!) rund 1,4 Mio. PageViews rund 8.000 PageViews/Tag rund 0,35 Mio. Visits rund 0,18 Mio. "Besucher")

Die Anzahl der Besucher-Zugriffe auf Schlüssel-URLs des IZ-Angebotes (Summe für den Zeitraum Februar - Dezember 2001) lag in der Spitzengruppe bei 3-8.000 PageViews pro URL.

Die monatlichen Spitzenwerte lagen bei 4-6.000 Benutzer-Zugriffe/URL/Monat für Index-Dateien und bis zu 2.300 Benutzer-Zugriffe/URL/Monat für spezielle, aktuelle Inhalts-URLs (z.B. FokusPlus Amerika).

Über 80 % der Client-IP-Adressen ("Besucher") wurden im Auswertungszeitraum nur einmal registriert.

Das Interesse der verschiedenen Suchmaschinen-Anbieter am GESIS-Angebot (gemessen in PageViews (PV)) war in den 11 Monaten des Untersuchungszeitraumes sehr unterschiedlich und reichte von 300.000 PV (Google) bis zu 1.000 PV (Lycos) pro Untersuchungszeitraum.

Eine AdClick-Analyse zeigt ergänzend, dass in diesem Zeitraum z.B. rund 52.000 Besucher-PageViews auf das GESIS-Angebot als Links aus Google-Suchen generiert wurden.

In Kap. 6 wird abschließend das ab 2002 im IZ implementierte Produktionsverfahren zur Auswertung der WWW-Log-Dateien einschließlich der dabei verwendeten Skripte und Filter-Dateien beschrieben und dokumentiert.

Kap. 8 Anhang enthält die Liste der Dateien des Produktionsverfahrens, die in ihrer jeweils aktuellen Version auf Anforderung an Interessenten geliefert werden können, sowie die Kontakt-E-Mail-Adressen für Rückfragen und weitere Informationen.

8 Anhang

Folgende Dateien sind auf Anforderung lieferbar:

- Source-Code der Filter-Prozeduren,
- Filterliste Suchmaschinen,
- Filterliste GESIS-Nutzer,
- Filterliste URLs,
- diese Dokumentation.

Kontakte: mell@bonn.iz-soz.de
beier@berlin.iz-soz.de
kunz@berlin.iz-soz.de